REPORT RESUMES

ED 003 389 24
EXPLORATION OF NOVEL RESEARCH DESIGNS AND MEASUREMENT TECHNIQUES.
BY- CAMPBELL, DONALD T.
NORTHWESTERN UNIV., EVANSTON, ILL.
REPORT NUMBER CRP-C-998 PUB DATE 65
CONTRACT OEC-3-20-001
EDRS PRICE MF-$1.62 HC-$44.80 1120P.

DESCRIPTORS- *EDUCATIONAL RESEARCH, *MEDIA RESEARCH, *SOCIAL SCIENCES, *MEASUREMENT TECHNIQUES, EVALUATION, *DATA COLLECTION, RESEARCH PROJECTS, EXPERIMENTAL PROGRAMS, *RESEARCH METHODOLOGY, EVANSTON, ILLINOIS

THE GOAL OF THIS PROJECT WAS AN EXPLORATION OF NOVEL RESEARCH DESIGNS AND MEASUREMENT TECHNIQUES SUITABLE FOR EMPLOYMENT IN MEDIA RESEARCH, EDUCATIONAL RESEARCH, OR IN SOCIAL SCIENCES. TWENTY-THREE RESEARCH REPORTS TOGETHER WITH A COVERING MEMORANDUM, PROVIDED THE SUBSTANCE OF THE TECHNICAL REPORT. THE MEMORANDUM INCLUDED THE RATIONALE FOR THE DIRECTION OF THE PROGRAM ON QUASI-EXPERIMENTAL DESIGN AND AN ANNOTATED BIBLIOGRAPHY WHICH SERVED BOTH AS THE INTRODUCTION TO THE REPORT AND AS A TABLE OF CONTENTS. (HB)

ED003389

List of Items constituting the final report from Donald T. Campbell, Northwestern University, on Project C-998, Contract 3-20-001, Educational Media Branch, U. S. Office of Health, Education, and Welfare, entitled "Explorations of Novel Research Designs and Measurement Techniques."

"Final Technical Report," a covering memorandum describing the 23 following research reports which constitute the body of the final report.

1. "Quasi-Experimental Design"
2. "From Description to Experimentation: Interpreting Trends as Quasi-Experiments."
3. "An Exploration of Two Quasi-Experimental Designs: the Cross-Lagged Panel Correlation and the Multiple Time-Series."
4. "Causal Relations in Attitude Change as Demonstrated through the Cross-Lagged Panel Correlation."
5. "A Study of the Effect of Proximally Auto-correlated Error on Tests of Significance for the Interrupted Time-Series Quasi-Experimental Design."
6. "The Interrupted Time-Series as Quasi-Experiment: Three Tests of Significance. A Fortran Program for the C.D.C. 3400 Computer."
7. "Time Series Data in the Quasi-Experimental Analysis of the Connecticut Speeding Crackdown."
8. "A Multiwave Panel Analysis of the Effects of a Political Speaker Upon a Self-Selected Audience."
9. "Invited Therapy in an Archival or Institutional Records Setting: With Comments on the Problem of Turndowns."
10. "Evaluating an Experimental College Program with Institutional Records: An Interim Report."
11. "An Evaluation of Thorndike's Classic Studies of Formal Discipline."
12. "Administrative Experimentation, Institutional Records, and Nonreactive Measures."
13. "Other Measures: A Survey of Unconventional and Cooperation-Free Measures for Social Science."
14. "Seating Aggregation as an Index of Attitude."
15. "A Procedure for Estimating the Amount of Trait, Method, and Error Variance Attributable to a Measure."
16. "Method Factors in Multitrait-Multimethod Matrices: Multiplicative Rather Than Additive?"
17. "Direct Iterative Factoring in Matrices with Crossed Factor Classes."
18. "The Sense of Identity: Measurement and an Examination of Correlates."
19. "A Fortran IV Program for Evaluating Internal Consistency and Single Factoredness in Sets of Multilevel Attitude Items."
20. "On the Use of Both Pro and Con Items in Attitude Scales."
21. "Some Simplex Matrices Based on High School Grades and Absences."
22. "Zero-Point Scaling as Affected by Social Object, Scaling Method, and Context."
23. "Pattern Matching as an Essential in Distal Knowing."

C-998 I

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

ED003389

Final Technical Report
Contract 3-20-001
Project C-998
Educational Media Branch
Office of Education
U.S. Department of Health, Education, and Welfare
(Under Title VII of the National Defense Education Act)

EXPLORATION OF NOVEL RESEARCH DESIGNS AND MEASUREMENT TECHNIQUES

Donald T. Campbell
Principal Investigator

Northwestern University, Contractor
July 15, 1962 to September 30, 1965

II

The goal of this contract has been the exploration of novel research designs and measurement techniques suitable for employment in media research, educational research, or in social science more generally. The direction of novelty particularly intended was toward appropriateness in natural field situations, in contrast to artificial laboratory settings.

It is an unfortunate byproduct of our increasing emphasis upon scientific rigor that it has produced an increasing artificiality. More and more, the laboratories and the problems of our experiments are totally dissimilar from the situations to which we wish to apply our findings--a trend which seriously jeopardizes the validity of such generalization. The direction of innovation sought in this series of explorations is to move some of the rigor of the laboratory science into field situations, and to utilize an experimental orientation in interpreting data analysis settings which are only quasi-experimental.

This general field of endeavor can be divided into the two parts indicated in the title. On the one hand there is the problem of research designs appropriate to field situations, to the settings in which people work and live. The work in the present contract continues that described in "Factors Relevant to the Validity of Experiments in Social Settings" (Psychological Bulletin, 1957, 54, 297-312) and in "Experimental and Quasi-Experimental Designs for Research on Teaching (with J.C. Stanley, in N.L. Gage, ed., Handbook of Research on Teaching, Rand McNally, 1963, pp. 171-246).

The second aspect is the problem of the measurement or observational devices usable in social settings. In this work, there is a continuation of work on validity started in the paper "Convergent and discriminant validation by the multitrait-multimethod matrix" (with D.W. Fiske, Psychological Bulletin, 1959, 56, 81-105). There is also attention to the internal consistency problem. Most important, however, is an emphasis on "nonreactive" measures, measures which avoid the artificiality of questionnaires and tests, since the latter may as often create attitudes as measure them and may at very least create on the part of respondents a self-consciousness and awareness of being tested atypical of the conditions to which we would like to generalize our findings.

Back of these concrete methodological issues lies a particular perspective on the enterprise of science and knowing, a philosophy of science or psychology of science, if you will, as shown in "Methodological Suggestions from a Comparative Psychology of Knowledge Processes" (Inquiry, 1959, 2, 152-182). It was envisaged in the initial contract that further work at this level might also be undertaken.

The products are a series of research reports scattered along this broad front. As explorations, they have been opportunistic, located as collaborators, data, designs, and conceptual analyses, became available. While as in any exploratory process, some blind alleys, disappointing outcomes, and unfinished starts have been encountered, en toto, it is believed, the exploration has been profitable.

This final report, then, consists of this covering memorandum and the 23 separate reports which accompany it. The following paragraphs list these in the context of the major goals as just reviewed, serving thus as both introduction and table of contents.

RESEARCH DESIGNS

The program of quasi-experimental design is introduced by a report prepared for the forthcoming International Encyclopedia of the Social Sciences:

1. "Quasi-Experimental Design"

This paper will also appear in slightly modified form in the 1965 Proceedings of the Social Statistics Section, American Statistical Association, as it was used as the basis for a symposium presentation at the meetings in Philadelphia on September 9th. The contract has also been represented in the revision of another, longer, general survey:

2. "From Description to Experimentation: Interpreting Trends as Quasi-Experiments", in C.W. Harris, Ed., Problems in Measuring Change: Madison: University of Wisconsin Press, 1963.

One of the highly speculative experimental designs given detailed exploration in this contract has been the Cross-Lagged Panel Correlation. In work on this design we have encountered primarily new grounds for caution. The design remains, however, the most promising of the current crop of efforts to reject the old taboo against inferring causation from correlation. The following two papers report our major efforts along this line:

3. "An Exploration of Two Quasi-Experimental Designs: the Cross-Lagged Panel Correlation and the Multiple Time-Series," by

Richard M. Rozelle.

4. "Causal Relations in Attitude Change as Demonstrated through the Cross-Lagged Panel Correlation," by Richard M. Rozelle.

While it is not presented as a research report here, the principal investigator delivered a paper reviewing these and other explorations of the cross-lagged panel correlation in a contribution to a symposium entitled: "Quasi-Experimental Design in Educational Research" at the annual meeting of the American Psychological Association in Chicago on September 5, 1965. The other two empirical papers on that symposium also dealt with uses of the cross-lagged panel correlation, finding more encouraging results than we have encountered. If, as seems likely, these presentations receive joint publication, our presentation will of course contain acknowledgement to this contract.

The second specific design receiving major exploration is the Interrupted Time-Series Design, sometimes accompanied by the Multiple Time Series elaboration. This is probably the most promising of the little-used quasi-experimental designs. Two studies bear on the problem of appropriate tests of significance for this design, and offer a practical if not theoretical solution to the problem of proximally auto-correlated error characteristic of such data series:

5. "A Study of the Effect of Proximally Auto-correlated Error on Tests of Significance for the Interrupted Time-Series Quasi-Experimental Design," by Joyce Sween and Donald T. Campbell.

6. "The Interrupted Time-Series as Quasi-Experiment: Three Tests of Significance. A Fortran Program for the C.D.C. 3400 Computer," by Joyce Sween and Donald T. Campbell. Distributed by the Vogelback Computing Center, Northwestern University, as NUC0049 TIMEX.

The Time Series design has been used in three investigations. One of these is that by Rozelle in report #3, above. The most publishable is:

7. "Time Series Data in the Quasi-Experimental Analysis of the Connecticut Speeding Crackdown," by H. Lawrence Ross and Donald T. Campbell.

The problem of measuring the persuasive effects of messages upon an audience which voluntarily seeks out attendance knowing the attitudes of the speaker is one which has not yet been solved. The following essay reports on a heroic, if frustrating, effort to establish a periodic opinion survey in a college to provide data for a multiple time-series approach to the problem:

8. "A Multiwave Panel Analysis of the Effects of a Political Speaker Upon a Self-Selected Audience," by Thomas Mader and Carole R. Siegman.

One type of quasi-experimental design which will probably become more frequent and important in the future is one in which a genuine randomization occurs at the level of invitation to treatment, but where such a large proportion of those invited fail to accept the invitation and get the experimental treatment that the design can no longer be regarded as a "true" experiment. The situation may not be entirely hopeless, however, providing clear awareness of the threats to validity is maintained. Work on this design remains incomplete. The general problem is stated in the first of the relevant reports, and an interim statement of an effort to apply it is found in the second:

9. "Invited Therapy in an Archival or Institutional Records Setting: With Comments on the Problem of Turndowns."

10. "Evaluating an Experimental College Program with Institutional Records: An Interim Report," by Sally W. Cassidy, Paule Verdet, Richard H. Schell, and Donald T. Campbell.

Other reports in the design area are the following two:

11. "An Evaluation of Thorndike's Classic Studies of Formal Discipline," by Paul C. Rosenblatt.

12. "Administrative Experimentation, Institutional Records, and Nonreactive Measures."

MEASUREMENT TECHNIQUES

Perhaps the contract's major contribution to the measurement area is in the emphasis upon nonreactive measures, measures which are so much a part of the everyday background of the respondents, or are so inconspicuous that they do not initiate change processes nor atypical responses to experimental treatments. This has been of course a major emphasis in our present and previous discussions of experimental design. As the titles of some indicate, the experimental programs described in reports 3, 7, 9, 10, 11, and 12, above, have involved institutional records rather than tests, questionnaires, or interviews. The most important product along this line, however, has been the following methodological analysis and survey of the literature:

13. "Other Measures: A Survey of Unconventional and Cooperation-Free Measures for Social Science," by Eugene J. Webb, Donald T. Campbell, and Richard D. Schwartz.

This survey has been circulated widely among persons thought to be experts in this field, and has met with both specific suggestions for addition and generous praise. In a considerably revised form, with more illustrations, with, sad to say, fewer efforts at humor, with Lee B. Sechrest added as a fourth author, and with a new title, possibly Unobtrusive Measures, it will be published by Rand McNally early in

1965. While the only acknowledgment of financial support will be to this contract, it should be noted that no personal support was provided Webb, Schwartz, or Sechrest. The contract supplied only research assistants, typists, etc. In the opinions of some, this book will be the most important contribution of the contract. As explorations of the kinds of methods recommended in this volume, items 3, 7, and 10, listed above in the Design section, should be examined, and also:

14. "Seating Aggregation as an Index of Attitude," by Donald T. Campbell and William P. Wallace, with an appendix by William H. Kruskal.

This paper in revised form will appear in Sociometry early in 1965, with W.H. Kruskal listed as second author.

A second major emphasis in the measurement field has been upon developing and improving the multitrait-multimethod matrix technique. While no journal publications have as yet emerged, significant advances have been made, which will eventually result in publications bearing acknowledgment to this contract. The following reports illustrate these gains:

15. "A Procedure for Estimating the Amount of Trait, Method, and Error Variance Attributable to a Measure," by Leroy Wolins.

16. "Method Factors in Multitrait-Multimethod Matrices: Multiplicative Rather Than Additive?" by Donald T. Campbell and Edward J. O'Connell.

17. "Direct Iterative Factoring in Matrices with Crossed Factor Classes."

Because one of the great problems in applying summary statistics to actual instances of multitrait-multimethod matrices is large sampling error due to small numbers of cases, this contract supplied the funds necessary to extend the N from 100 to 700, plus typing and duplicating costs, in the following study, which also explores novel measurements in an uncharted personality domain:

18. "The Sense of Identity: Measurement and an Examination of Correlates," by Frank Baker.

The problem of internal consistency, of justifying the pooling of responses to many items in one total score, is a recurrent one in the social sciences. We believe we have the best solution to date in the following computer program, which utilizes among other indicators, a Lawley Maximum Likelihood Factor analysis, and an "index of single factoredness" comparing obtained residuals with those which would have been obtained by chance for a one-factored case with the given N and level of inter-item correlation:

19. "A Fortran IV Program for Evaluating Internal Consistency and Single Factoredness in Sets of Multilevel Attitude Items" by Donald G. Morrison, Donald T. Campbell, and Leroy Wolins, distributed by the Vogelback Computing Center, Northwestern University, as NUC0048 ATTANAL.

Also relevant to the problem of internal consistency are the following two items, the first of which deals with controlling irrelevant direction-of-wording factors, the second of which is a minor byproduct from report No. 3, of possible help to those working on internal consistency from the point of view of Guttman's simplex factor analysis:

20. "On the Use of Both Pro and Con Items in Attitude Scales."

21. "Some Simplex Matrices Based on High School Grades and Absences," by Richard M. Rozelle and Donald T. Campbell.

A contribution to the effort to anchor attitude scales or judgment procedures in some absolute zero is contained in the following, which in shortened form will soon appear in the Journal of Social and Personality Psychology:

22. "Zero-Point Scaling as Affected by Social Object, Scaling Method, and Context," by Jack M. Hicks and Donald T. Campbell.

STRATEGY OF SCIENCE

In the plan of operation of the initial contract, as the last item in a list of twelve possible "specific activities" to be carried out under the contract, the following occurred:

> Problems in the theory of inference: the general problem of the relationship between data and theory; the relationship between measurement operations and the definition of theoretical constructs; the concept of "validation" or "proof" of a theory in scientific practice; contributions of the psychology of knowledge processes to the strategy of science.

Within this fascinating and difficult agenda, only one paper has been prepared. I am happy to report that I judge it one of my best essays:

23. "Pattern Matching as an Essential in Distal Knowing," in K.R. Hammond, Ed. Egon Brunswik's Psychology. New York: Holt, Rinehart & Winston, late 1965 or early 1966.

These twenty three research reports, together with this coverinng memorandum, are submitted as the final technical report on Contract 3-20-001, Project C998, with the Media Research Branch of the U.S. Office of Education.

Donald T. Campbell

Donald T. Campbell
Professor of Psychology
Northwestern University
Evanston, Illinois

October 30, 1965

Prepared for the *International Encyclopedia of the Social Sciences*, David L. Sills, editor, to be published ca. 1966 by The Macmillan Company, The Free Press of Glencoe, and The Crowell-Collier Publishing Company

Quasi-Experimental Design[1]

Donald T. Campbell
Northwestern University

This phrase refers to the application of an experimental mode of analysis and interpretation to bodies of data not meeting the full requirements of experimental control. The settings to which it is appropriate are those of experimentation in social settings, including planned interventions such as specific communications, persuasive efforts, changes in conditions and policies, efforts at social remediation, etc. Unplanned conditions and events may also be analyzed in this way where an exogenous variable has such discreteness and abruptness as to make appropriate its consideration as an experimental treatment applied at a specific point in time to a specific population. When properly done, when attention is given to the specific implications of the specific weaknesses of the design in question, quasi-experimental analysis can provide a valuable extension of the experimental method.

While efforts to interpret field data as experiments go back much farther, the first prominent methodology of this kind in the social sciences was Chapin's Ex Post Facto Experiment (Chapin & Queen, 1937;

---

[1] The preparation of this review was supported in part by Project C-998, Contract 3-20-001, with the Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

Chapin, 1955; Greenwood, 1945), although it should be noted that due to the failure to control regression artifacts, this mode of analysis is no longer regarded as acceptable. The American Soldier volumes (Stoufer et al., 1949) provide prominent analyses of the effects of specific military experiences, where it is implausible that differences in selection explain the results. Thorndike's efforts to demonstrate the effects of specific course work upon other intellectual achievements provide an excellent early model (e.g., Thorndike & Woodworth, 1901; Thorndike & Ruger, 1923). Extensive analysis and review of this literature are provided elsewhere (Campbell, 1957; 1963; Campbell & Stanley, 1963) and serves as the basis for the present abbreviated presentation.

The core requirement of a "true" experiment lies in the experimenter's ability to apply experimental treatments in complete independence of the prior states of the materials (persons, etc.) under study. This independence makes resulting differences interpretable as effects of the differences in treatment. In the social sciences this independence of prior status is assured by randomization in assignments to treatments. Experiments meeting these requirements, and thus representing "true" experiments, are much more possible in the social sciences than is generally realized. Wherever, for example, the treatments can be applied to individuals or small units (such as precincts or classrooms) without the respondents being aware of experimentation or that other units are getting different treatments, very elegant experimental control can be achieved. An increased acceptance by administrators of randomization as the democratic method of allocating scarce resources (be these new housing, therapy, or fellowships) will make possible field experimentation in many settings. Where innovations are to be introduced throughout a social system, and where the introduction cannot in any event be simultaneous, a use of randomization in the staging can provide an experimental comparison of the new and the old, using the groups receiving the delayed introduction as controls. Nothing in this article should be interpreted as minimizing the importance of increasing the use of true experimentation. However, where true experimental design with random assignment of persons to treatments is not possible, due to ethical considerations, lack of power or feasibility, application of quasi-experimental analysis has much to offer.

The social sciences must do the best they can with the possibilities open to them. Inferences must frequently be made from data lacking complete control. Too often a scientist trained in experimental method rejects out of hand any research in which complete control is lacking. Yet in practice no experiment is perfectly executed, and the practicing scientist overlooks those imperfections which seem to him to offer no plausible rival explanation of the results. In the light of modern philosophies of science, no experiment ever proves a theory, it merely probes it. Seeming proof results from that condition in which there is no available plausible rival hypothesis to explain the data. The general program of quasi-experimental analysis is to specify and examine those plausible rival explanations of the results which are provided by the uncontrolled variables. A failure of control which does not in fact provide a plausible rival interpretation is not regarded as invalidating.

It is well to remember that we do make assured causal inferences in many settings not involving randomization: (The earthquake caused the brick building to crumble; the automobile crashing into it caused the telephone pole to break; the language patterns of the older models and mentors caused this child to speak English rather than Kwakiutl; etc.) While these are all potentially erroneous inferences, they are of the same type as experimental inferences. We are confident that were we to intrude experimentally, we could confirm the causal laws involved. Yet they have been made assuredly by a nonexperimenting observer. This assurance is due to the effective absence of other plausible causes. Consider the inference as to crashing auto and the telephone pole: we rule out combinations of termites and wind because the other implications of these theories (e.g., termite tunnels and debris in the wood, wind records at nearby weather stations) do not occur. Spontaneous splintering of the pole by happenstance coincident with the auto's onset does not impress us as a rival, nor would it explain the damage to the car, etc. Analogously in quasi-experimental analysis, tentative causal interpretation of data may be made where the interpretation in question squares with the data and where other rival interpretations have been rendered implausible.

For the evaluation of data series as quasi-experiments, a set of twelve frequent threats to validity have been developed. These may be regarded as the important classes of frequently plausible rival hypotheses which good research design seeks to rule out. All will be presented briefly even though not all are employed in the evaluation of the designs used illustratively here.

Fundamental to this listing is a distinction between _internal validity_ and _external validity_. _Internal validity_ is the basic minimum without which any experiment is uninterpretable: did in fact the experimental treatments make a difference in this specific experimental instance? _External validity_ asks the question of _generalizability_: to what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds, in that features increasing one may jeopardize the other. While internal validity is the sine qua non, and while the question of external validity, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.

Relevant to internal validity, are eight different classes of extraneous variables which, if not controlled in the experimental design, might produce effects mistaken for the effect of the experimental treatment. These are: 1. History: the other specific events occurring between a first and second measurement in addition to the experimental variable. 2. Maturation: processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing tireder and the like. 3. Testing: the effects of taking a test upon the scores of a second testing. 4. Instrumentation: in which changes in the calibration of a measuring instrument or changes in the observers of scorers used may produce changes in the obtained measurements. 5. Statistical regression: operating where groups have been selected on the basis of their extreme scores. 6. Selection: biases resulting in differential recruitment of respondents for the comparison groups. 7. Experimental mortality: the differential loss of respondents from the comparison groups. 8. Selection-maturation interaction: in certain of the multiple-group quasi-experimental designs, such as the nonequivalent control group design, such interaction is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

Factors jeopardizing external validity or representativeness are: 9. The reactive or interaction effect of testing, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected. 10. Interaction effects between selection bias and the experimental variable. 11. Reactive effects of experimental arrangements, which would preclude generalization about the effect of the experimental variable for persons being exposed to it in nonexperimental settings. 12. Multiple-treatment interference, a problem wherever multiple treatments are applied to the same respondents, and a particular problem for one-group designs involving equivalent time-samples or equivalent materials samples.

Perhaps the simplest quasi-experimental design is the One-Group Pretest-Posttest Design, $\underline{O}_1$ $\underline{X}$ $\underline{O}_2$ (where $\underline{O}$ represents measurement or observation $\underline{X}$ the experimental treatment). This common design patently leaves uncontrolled the internal validity threats of History, Maturation, Testing, Instrumentation, and, if selected as extreme on $\underline{O}_1$, Regression. There may be situations in which the analyst could decide that none of these represented plausible rival hypotheses in his setting: A log of other possible change-agents might provide no plausible ones, the measurement in question might be nonreactive (Campbell, 1957), the time span too short for maturation, too spaced for fatigue, etc. However, the sources of invalidity are so numerous that a more powerful quasi-experimental design would be preferred. Several of these can be constructed by adding features to this simple one. The interrupted Time-Series Experiment utilizes a series of measurements providing multiple pretests and posttests, e.g.: $\underline{O}_1 \quad \underline{O}_2 \quad \underline{O}_3 \quad \underline{O}_4\,\underline{X} \quad \underline{O}_5 \quad \underline{O}_6 \quad \underline{O}_7 \quad \underline{O}_8$ If in this series, $\underline{O}_4 - \underline{O}_5$ shows a rise greater than found elsewhere, then Maturation, Testing, and Regression are no longer plausible, in that they would predict equal or greater rises for $O_1 - O_2$, etc. Instrumentation may well be controlled too, although in institutional settings, a change of administration policy is often accompanied by a change in record-keeping standards. Observers and participants may be focused on the occurrence of $\underline{X}$, and may speciously change rating standards, etc. History remains the major threat, although in many settings it would not offer a plausible rival interpretation. If one had available a parallel time series from a group not receiving the experimental treatment, but exposed to same extraneous sources of influence, and if this control time series failed to show the exceptional jump from $\underline{O}_4$ to $\underline{O}_5$, then the plausibility of History as a rival interpretation would be greatly reduced. We may call this the Multiple Time-Series Design.

Another way of improving The One-Group Pretest-Posttest Design is to add a "Nonequivalent Control Group." (Were the control group to be randomly assigned from the same population as the experimental group, we would, of course, have a true, not quasi, experimental design.) Depending on the similarities of setting and attributes, if the nonequivalent control group fails to show a gain manifest in the experimental group, then History, Maturation, Testing, and Instrumentation are controlled. In this popular design, the frequent effort to "correct" for the lack of perfect equivalence by matching on pretest scores is absolutely wrong (e.g., Thorndike, 1942; Hovland et al., 1949; Campbell & Clayton, 1961), as it introduces a regression artifact. Instead, one should live with any initial pretest differences, using analysis of covariance, gain scores, or graphic presentation. Remaining uncontrolled is the Selection-Maturation Interaction, i.e., the possibility that the experimental group differed from the control group not only in initial level, but also in its autonomous maturation rate. In experiments on psychotherapy and on the effects of specific coursework this is a very serious rival. Note that it can be rendered implausible by use of a time series of pretests for both groups thus moving again to the Multiple Time-Series Design.

There is not space here to present adequately even these four quasi-experimental designs but perhaps the strategy of adding specific observations and analyses to check on specific threats to validity has been illustrated. This is carried to an extreme in the Recurrent Institutional Cycle Design (Campbell & McCormack, 1957; Campbell & Stanley, 1963), in which longitudinal and cross-sectional measurements are combined with still other analyses to assess the impact of indoctrination procedures, etc., through exploiting the fact that essentially similar treatments are being given to new entrants year after year or cycle after cycle. Other quasi-experimental designs covered in Campbell and Stanley (1963) include two more single-group designs (The Equivalent Time-Samples Design and The Equivalent Materials Design), Counterbalanced or Rotational Designs, Separate Sample Pretest-Posttest Designs, Regression-Discontinuity Analysis, The Panel Impact Design (see also Campbell & Clayton, 1961), and the Cross-Lagged Panel Correlation which is related to Lazarsfeld's Sixteen-Fold Table (see especially Campbell, 1963).

Related to the program of quasi-experimental analysis are those efforts to achieve causal inference from correlational data. Note that while correlation does not prove causation, most causal hypotheses imply specific correlations, and examination of these thus probes, tests, or edits, the causal hypothesis. Further, as Simon and Blalock have emphasized (e.g., Blalock, 1960) certain causal models specify uneven patterns of correlation. Thus the A → B → C model implies that r_{AC} be smaller than r_{AB} or r_{BC}. However, their use of partial correlations or the use of Wright's (1920) path analysis are rejected as tests of the model because of the requirement that the "cause" be totally represented in the "effect." In the social sciences it will never be plausible that the "cause" has been measured without unique error and that it also totally lacks unique systematic variance not shared with the "effect." More appropriate would be Lawley's (1940) test of the hypothesis of single-factoredness. Only if single-factoredness can be rejected would the causal model as represented by its predicted uneven correlation pattern be the preferred interpretation.

References

Blalock, H.M. 1960 *Social statistics*. New York: McGraw-Hill.

Campbell, D. T. 1957 Factors relevant to validity of experiments in social settings. *Psychological Bulletin* 54:297-312.

Campbell, D. T. 1963 From description to experimentation: Interpreting trends as quasi-experiments. Pages 212-242 in Harris, C. W. (editor), *Problems in measuring change*. Madison, Wis.: Univ. Wisconsin Press.

Campbell, D T.; and Clayton, K.N. 1961 Avoiding regression effects in panel studies of communication imp[illegible]t. *Studies in Public Communication* No. 3, 99-118.

Campbell, D.T.; and McCormack, Thelma H. 1957 Military experience and attitudes toward authority. *American Journal of Sociology* 62:482-490.

Campbell, D.T.; and Stanley, J.S. 1963 Experimental and quasi-experimental designs for research on teaching. Pages 171-246 in Gage, N L. (editor), *Handbook of research on teaching*. Chicago: Rand McNally.

Chapin, F.S. 1955 *Experimental designs in sociological research*. New York: Harper. (Rev. ed.)

Chapin, F.S.; and Queen, S.A. 1937 *Research memorandum on social work in the depression*.

Greenwood, E. 1945 *Experimental sociology: A study in method*. New York: King's Crown Press.

Hovland, C.I.; Lumsdaine, A.A.; and Sheffield, F.D. 1949 *Experiments on mass communication*. Princeton, N.J.: Princeton Univ. Press.

Lawley, C. N. 1940 The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh* 60:64-82

Stouffer, S. S. (editor) 1949 *The American soldier*. Princeton, N.J.: Princeton Univ. Press. Vols. I and II.

Thorndike, R L. 1942 Regression fallacies in the matched groups experiment. *Psychometrika* 7:85-102.

Thorndike, E. L.; and Ruger, G.J. 1923 The effect of first-year Latin upon knowledge of English words of Latin derivation. *School and Society* 81:260-270, 417-418.

Thorndike, E. L.; and Woodworth, R.S. 1901 The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review* 8:247-261, 384-395, 553-564.

Wright, S. 1920 Correlation and causation. *Journal of Agricultural Research* 20:557-585.

Reprinted from Chester W. Harris, Ed. Problems in Measuring Change
Madison: The University of Wisconsin Press, 1963

Twelve

From Description to Experimentation: Interpreting Trends as Quasi-Experiments

DONALD T. CAMPBELL

The preparation of this paper has been made possible by the author's participation in Northwestern University's Carnegie Corporation supported program in psychology and education. Revision of it has been supported by U.S. Office of Education Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

Scientist as intruder or scientist as unintrusive observer: These two models compete for the loyalties of social scientists. In this conference on the measurement of change we have primarily focused upon our role as describers of processes that would be going on even had we not been present. We have examined the methodology of asking: "Did a change take place here?" "What attributes changed?" "What things tend to change together?" "What is the general, normal, change or growth pattern for this animal, this person, this group, this institution?" "How do change patterns vary as concomitants of social settings?" The problems encountered at this level are so formidable that it may seem gratuitous to ask a still more ambitious one. So frequently do data processing procedures mislead us in answering the question "Did a change occur?" that we should perhaps hesitate before jumping to ask "What caused the change?"

Yet in the literature, the methodology of which we are seeking to improve, this reticence is usually lacking. Few describe change without attempting to explain it. Most undertake the research assuming in advance that they know the causal agent, asking only the question of the degree of efficacy: be it of genetic patterns of maturation, Vassar College, military experience, psychotherapy, remedial reading, or what not. Correspondingly, our meta-science of methodology should stay with our clients over this interpretative hurdle too.

The present author accepts "true" experimentation as the optimal model for the interpretation of change in causal terms. With regard to the models of scientist as intruder (experimenter) versus scientist as unintrusive observer, I clearly prefer the first where it is possible, and would utilize the second only as a substitute where intrusion is not possible. Experimental design in the Fisher tradition is the optimal setting for the study of interpretable change. The magic of randomization is that it attenuates the causal threads of the past as they might codetermine both exposure to the treatment and gain scores. Randomization renders implausible innumerable rival explanations of the observed change by cutting the lawful relationships which in the natural setting would determine which person gets which treatment. This provides the persuasive "causal" interpretations, made possible by experiments involving randomization. Where he can, the social scientist interested in studying change should so intrude.

Yet we make assured causal inferences in many settings not involving randomization: The earthquake caused the brick building to crumble; the car's crashing into it caused the telephone pole to break; the language patterns of the older models and mentors caused this child to speak English instead of Kwakiutl. These are all potentially erroneous inferences, unproven, unsupported by deductive certainty. They are of the same type as experimental inferences, in that we are confident that were we to intrude experimentally, we could confirm the causal laws involved. Yet they are inferences available to the unintrusive observer. Cannot the social scientist as observer make some inferences as to sources of change on comparable grounds? The program of attempting quasi-experimental interpretation of observational sequences addresses itself to this problem.

***Some slogans*: Experiments *probe* theory, but do not *prove* theory. Any data that probe theory are worth looking at. The only uncontrolled variables that invalidate an experiment are those to which we are willing to attribute the status of empirical laws comparable in standing to the law under investigation. When we have an evidence of change and wish to interpret it causally, the only relevant threats to that interpretation are other plausible, probable, causal explanations.**

Consider the inference as to crashing car and telephone pole. Our assurance of inference is due to the total absence of other possible causes. Combinations of termites and wind we have ruled out because the other data implications of these theories (e.g., termite tunnels and debris in the wood and wind records in nearby weather stations) do not occur. Spontaneous splintering of poles by happenstance coincident with the auto's onset does not impress us as a rival, nor would it explain the damage to

the car. Analogously, in the quasi-experimental program, field observations will be given causal interpretation where (*a*) the interpretation in question squares with the data, and (*b*) other plausible rival hypotheses have been ruled out.

Does this let down the bars and give approval to the descriptive journalism which so often passes for science in our fields? I think not, if we vigorously attend to the specific plausible rival hypotheses appropriate to each situation. In particular, I find the plausible rival hypothesis of regression (R. L. Thorndike, 1942; Campbell and Clayton, 1961) so compelling as to rule out totally these well-established customs in the reported observations of natural change: Chapin's (1955) "ex-post facto experiments" in all forms; matching as a substitute for randomization in providing pretreatment equivalence of groups; sociology's "qualifier," "elaboration," "turnover table," and other complex cross-classifications with percentages for comparison computed to part-sample bases; partial or multiple correlation when used to reject the hypothesis of cosymptom in favor of the hypothesis of distinctive determinant; and the division of pretest scores into extreme subgroups for the assessment of differential rates of change as a function of initial level.

Check-list of factors jeopardizing internal and external validity

For the evaluation of data series as experiments or quasi-experiments, a set of twelve frequent threats to validity have been developed (Campbell, 1957; Campbell and Stanley, 1963). These may be regarded as the important classes of frequently plausible rival hypotheses which good research design seeks to rule out. While not all will be employed in the evaluation of the designs treated here, all will be presented briefly so as to provide continuity with the previous presentations.

Fundamental to this listing is a distinction between *internal validity* and *external validity*. *Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental instance? *External validity* asks the question of *generalizability*: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds, in that features increasing one may jeopardize the other. While *internal validity* is the *sine qua non*, and while the question of *external validity*, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.

Relevant to *internal validity*, there are eight different classes of extraneous variables which, if not controlled in the experimental design, might produce effects confounded with the effect of the experimental stimulus. These are:

1. *History:* the other specific events occurring between the first and second measurement in addition to the experimental variable.
2. *Maturation:* processes within the respondents operating as a function of the passage of time *per se* (not specific to the particular events), including growing older, growing hungrier, growing more tired, and the like.
3. *Testing:* the effects of taking a test upon the scores of a second testing.
4. *Instrumentation:* changes in the calibration of a measuring instrument or changes in the observers or scorers which may produce changes in the obtained measurements.
5. *Statistical regression:* regression operating when groups have been selected on the basis of their extreme scores.
6. *Selection:* biases resulting in differential recruitment of respondents for the comparison groups.
7. *Experimental mortality:* the differential loss of respondents from the comparison groups.
8. *Selection-maturation interaction*, etc.: in certain of the multiple-group quasi-experimental designs, such as the nonequivalent control-group design, such interaction is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

The factors jeopardizing *external validity* or *representativeness* are:

9. The *reactive* or *interaction effect* of *testing*, in which a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected.
10. *Interaction* effects between *selection* bias and the *experimental variable*.
11. *Reactive effects of experimental arrangements*, which would preclude generalization about the effect of the experimental variable for persons being exposed to it in nonexperimental settings.
12. *Multiple-treatment interference*, a problem wherever multiple treatments are applied to the same respondents, and a particular problem for one-group designs involving equivalent time-samples or equivalent materials samples.

In the presentation of the experimental designs, a uniform code and graphic representation will be employed to epitomize most, if not all, of their distinctive features. An X will represent the exposure of a group to an experimental variable or event, the effects of which are to be measured; O will refer to some process of observation or measurement; the X's and O's in a given row are applied to the same specific persons. The left-to-right dimension indicates the temporal order, with X's and O's vertical to one another being simultaneous. These are the simplest symbols needed to designate the different designs as they are usually presented. To make the important distinction between the "true" experimental control-group design and the nonequivalent control-group design, a symbol R indicating random assignment to separate treatment groups is necessary. This randomization is conceived to be a process occurring at a specific time, and is the all-purpose procedure for achieving pretreatment equality of groups, within known statistical limits. Along with this goes another graphic convention, in that parallel rows unseparated by dashes represent comparison groups equated by randomization, while those separated by a dashed line represent comparison groups not equated by random assignment.

Table 12.1 provides a check-off of the designs discussed against these threats to validity. In this table, and in the discussion, the designs for convenience are numbered as in Campbell and Stanley (1963). In this paper the main emphasis is upon the two time-series designs and the cross-lagged panel correlation. The two more traditional quasi-experimental designs are presented as background argument against which to contrast the others.

The one-group pretest–post-test design

The simplest two-stage measure of change is also one of the commonest experimental designs, although so weak when used in the social sciences as not really to justify being called an experiment. It can be diagramed as follows:

$O_1 \; X \; O_2$ One-group pretest–post-test design.

One of the uncontrolled rival hypotheses for this design is *history*. Between O_1 and O_2 many other potentially change-producing events may have occurred in addition to the experimenter's X. If the pretest (O_1) and the post-test (O_2) are on different days, then the events in between may have caused the difference. To become a *plausible* rival hypothesis, such an event should have occurred to most of the subjects in the group under study, as in some other class-period or via a widely disseminated news story. In Collier's (1944) classroom study, while his students were reading Nazi propaganda materials, France fell; the obtained attitude

TABLE 12.1

Check-Off of Sources of Invalidity

	Internal								External			
Designs	History	Maturation	Testing	Instrumentation	Regression	Selection	Mortality	Interaction of selection and maturation, etc.	Interaction of testing and X	Interaction of selection and X	Reactive arrangements	Multiple-X interference
2. One-group pretest-post-test design $O \; X \; O$	–	–	–	–	?	+	+	–	–	–	?	
7. Time-series $O \; O \; O \; X \; O \; O \; O$	–	+	+	?	+	+	+	+	–	?	?	
10. Nonequivalent control-group design $O \; X \; O$ - - - - - $O \quad O$	+	+	+	+	?	+	+	–	–	?	?	
14. Multiple time-series $O \; O \; O \; X \; O \; O \; O$ - - - - - - - - - - - $O \; O \; O \quad O \; O \; O$	+	+	+	+	+	+	+	+	–	–	?	
17. Cross-lagged panel r X_1X_2 X O_1O_2	?	?	?	+	+	+	+	–	–			

Note: A minus indicates a definite weakness, a plus indicates the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates not relevant.

changes seemed more likely to be the result of this event than of the propaganda itself. *History* becomes a more plausible rival explanation of change the longer the O_1–O_2 time lapse, and might be regarded as a trivial problem in an experiment completed within a one-hour or two-hour period, although even here, extraneous sources, such as laughter, distracting events, etc., are to be looked for. Relevant to the variable *history* is the feature of *experimental isolation*, which can so nearly be achieved in many physical science laboratories as to render the one-group pretest–post-test design acceptable for much of their research. Such effective experimental isolation can almost never be assumed in social science. For these reasons a minus has been entered in Table 12.1 under *History*. We will classify with *history* possible effects of season or of institutional-event schedule, although these might also be placed with *maturation*. Thus optimism might vary with seasons and anxiety with the semester examination schedule (Windle, 1954; Crook, 1937). Such effects might produce an O_1–O_2 change confusable with the effect of X.

A second rival variable, or class of variables, is designated *maturation*. This term is used here to cover all of those biological or psychological processes which systematically vary with the passage of time, independent of specific external events. Thus between O_1 and O_2 the students may have grown older, hungrier, more tired, more bored, and the obtained difference may reflect this process rather than X. In social remediation or psychotherapy which focuses on exceptionally disadvantaged persons, a process of "spontaneous remission," analogous to wound healing, may be mistaken for the specific effect of a remedial X. (Needless to say, such a remission is not regarded as "spontaneous" in any causal sense, but rather represents the cumulative effects of learning processes and environmental pressures of the total daily experience, which would be operating even if no X had been introduced.)

A third confounded rival explanation is the effect of *testing*, the effect of the pretest itself. For achievement and intelligence tests it is usually the case that students taking the test for a second time, or taking an alternate form of the test, do better than those taking the test for the first time (Anastasi, 1958, pp. 190–191). These effects, as much as three to five IQ points on the average for naive test-takers, occur without any instruction as to scores or items missed on the first test. For personality tests, a similar effect is noted, with second tests showing in general better adjustment, although occasionally a highly significant effect in the opposite direction is found (Windle, 1954). For attitudes toward minority groups a second test may show more prejudice, although the evidence is very slight (Rankin and Campbell, 1955). Obviously, conditions of anonymity increased awareness of what answer is socially approved, and the like, all would have a bearing on the direction of the result. For prejudice items under conditions of anonymity, the adaptation level created by the hostile statements presented may shift the student's expectations as to what kinds of attitudes are tolerable in the direction of greater hostility. In a signed personality or adjustment inventory, the initial administration partakes of a problem-solving situation in which the student attempts to discover the disguised purpose of the test. Having done this, or having talked with his friends about their answers to some of the bizarre items, he knows better how to present himself acceptably the second time.

With the introduction of the problem of test effects comes a distinction among potential measures as to their *reactivity*. This is a most important theme in my overall program, accompanied by a general exhortation to use *nonreactive* measures wherever possible. It has long been a truism in the social sciences that the process of measuring may change that which is being measured. The test-retest gain would be one important aspect of such change, a threat to *internal* validity. The interaction of testing and X, which could occur without any main effect of testing, represents a serious threat to *external* validity. Such reactive effects can be expected whenever the testing process is in itself a stimulus to change rather than a passive record of behavior. Thus in an experiment on therapy for weight control, the initial weigh-in might in itself be a stimulus to weight reduction, even without the therapeutic treatment. Similarly, placing observers in the classroom to observe the teacher's pretraining human relations skills may in itself change the teacher's mode of discipline. Placing a microphone on the desk may change the group interaction pattern. In general, the more novel and motivating the test device, the more reactive one can expect it to be.

Instrumentation or "instrument decay" (Campbell, 1957) is the term used to indicate a fourth uncontrolled rival hypothesis. This term refers to autonomous changes in the measuring instrument which might account for an O_1–O_2 difference. These changes would be analogous to the stretching or fatiguing of a spring scales, or condensation in a cloud chamber. Where human observers are used to provide O_1 and O_2, processes of learning, fatiguing, etc., within the observers will produce O_1–O_2 differences. If essays are being graded, the grading standards may shift between O_1 and O_2 (suggesting the control technique of shuffling the O_1 and O_2 essays together and having them graded without knowledge of which came first). If classroom participation is being observed, then the observers may be more skillful or more blasé on the second occasion. If parents are being interviewed, the interviewer's familiarity with the interview schedule and with the particular parents may produce shifts. Similarly, a change in interviewers or observers between O_1 and O_2 could cause a difference.

A fifth confounded variable in some instances of the one-group pretest–post-test design is statistical *regression*. If, for example, in a remediation experiment, students are picked for a special experimental treatment because they are particularly poor on an achievement test (which becomes for them the O_1), then on a subsequent testing using a parallel form, or repeating the same test, O_2 for this group will almost surely average higher than did O_1. This dependable result is not due to any genuine effect of X or to any test-retest practice effect. It is rather a tautological aspect of the imperfect correlation between O_1 and O_2. Because of the great attention this or related problems have received at this conference, I shall not belabor this point here.

The interrupted time-series experiment

The essence of this experimental design is the presence of a periodic measurement process on some group or individual and the introduction of an experimental change into this time-series of measurements, the results of which are indicated by a discontinuity in the measurements recorded in the time-series. It can be diagramed thus:

$O_1\ O_2\ O_3\ O_4\ X\ O_5\ O_6\ O_7\ O_8$ Time-series design.

This experimental design might be regarded as typical of much of the classical, 19th century experimentation in the physical sciences and in biology, typified by the polygraph record for extended periods before and after a major change of conditions for a single specimen. Thus this experimental design is frequently regarded as valid in the more successful sciences even though it has no accepted status in the enumerations of available experimental designs in the social sciences. There are good reasons for this differential status, and a careful consideration of them will provide a better understanding of the conditions under which the design might meaningfully be employed by social scientists in the absence of the possibility of achieving more thorough experimental control.

Figure 12.1 indicates some possible outcome patterns for time-series into which an experimental alteration had been introduced as indicated by the vertical line X. For purposes of discussion it will be assumed that one will be tempted to infer that X had some effect in time-series with outcomes such as A and B and possibly in time-series such as C and D, but that one would not be tempted to infer an effect in time-series such as E, F, and G, even were the jump in values from O_4 to O_5 as great and as statistically stable as were the O_4 to O_5 differences in A and B, for example.

While the problem of statistical tests will be postponed for a few paragraphs, it is assumed that the problem of internal validity boils down to the question of plausible competing hypotheses that offer likely alternate explanations to the shift in the time-series other than the effect of X. A tentative check-off of the controls provided by this experiment under these optimal conditions of outcome is provided in Table 12.1. The strengths of the time-series design are most apparent in contrast with Design 2, to which it has a superficial similarity in lacking a control group and in using before and after measures.

When the list of problems of internal validity in Table 12.1 is scanned, it is immediately apparent that failure to control *history* is the most important weakness of Design 7. That is, the rival hypothesis exists that not X but some more or less simultaneous event produced the shift. It is upon the plausibility of ruling out such extraneous stimuli that credence to the interpretation of this experiment in any given instance must rest.

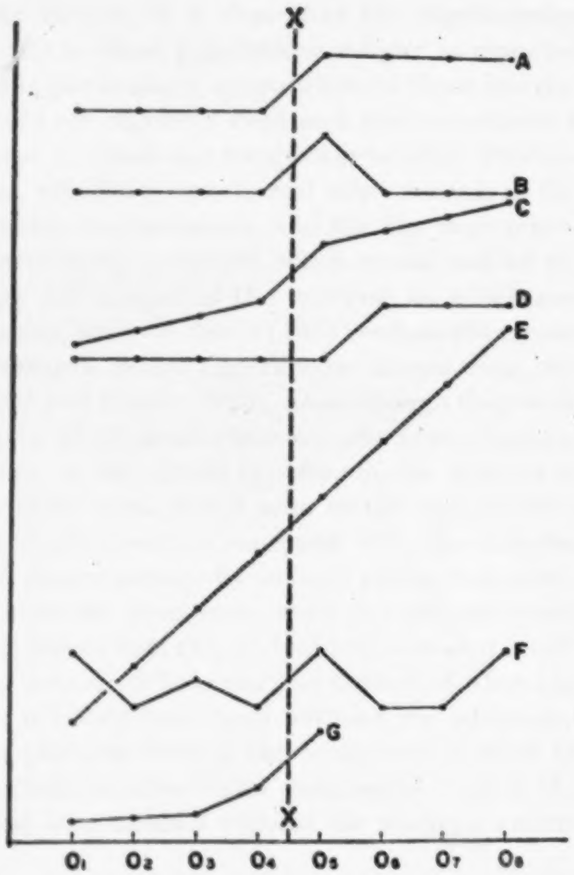


Fig. 12.1.—Some possible outcome patterns from the introduction of an experimental variable at point X into a time-series of measurements, O_1–O_8. (Except for D, the O_4–O_5 gain is the same for all time-series, while the legitimacy of inferring an effect varies widely, being strongest in A and B, and totally unjustified in E, F, and G.)

In many social science situations in which this design might be used, as in a factory or ship at sea, the experimenter might plausibly claim experimental isolation in the sense that he was aware of the possible rival events that might cause such a change and could assert that they did not occur in such a pattern as to provide an explanation of the results. It is on grounds such as these that Holmberg (1954) can refer to his experiences among the Siriono Indians as an experiment in culture change. Here was a tribe so isolated from contact with Europeans or other Indians that the anthropologist living among them for extended periods of time might well claim to be aware of the total impingement of novel stimulation. Observing as he was on a day-to-day basis, he might plausibly interpret in experimental terms the effects of his introduction of firearms and new crops. Insofar as his observations are to be fitted into an experimental paradigm at all, they would fit in best with the time-series design.

Among the other extraneous variables which might for convenience be put into *history* are the effects of weather and the effects of season. Experiments of this type are apt to extend over time periods that involve seasonal changes. In studies of worker output, the seasonal fluctuations in temperature and in the amount and quality of illumination available may be confounded with the introduction of the experimental change. Perhaps best also included under *history* would be periodical shifts in the time-series related to institutional customs such as the weekly work cycles, pay period cycles, and the like. The observational series should be such as to hold such known cycles constant or else be long enough to include several such cycles in their entirety.

To continue with the check-off factors to be controlled, *maturation* and *testing* are ruled out on the grounds that if the outcome is of such a nature as found in the illustrations *A* and *B* of Figure 12.1, *maturation* and *testing* do not provide plausible rival hypotheses to explain a shift occurring between O_4 and O_5 which did not occur in the previous time periods under observation. Had one only the observations at O_4 and O_5, as in Design 2, this means of rendering implausible *maturation* and test-retest effects would be lacking. Herein lies the great advantage of this design over Design 2. In a similar way, many hypotheses introducing changes in *instrumentation* would lack a specific rationale for expecting the instrument decay to occur on this particular occasion as opposed to earlier ones. However, a question mark has been entered at this point in Table 12.1 to call attention to several types of situations in which a change in the calibration of the measurement device could be misinterpreted as the effect of the experimental change. If the measurement procedure involves the judgments of human observers who are aware of the experimental plan, pseudo-confirmation of the hypothesis can occur due to the observer's expectations. Thus the experimental change of instituting a new police commissioner may produce a change in the reporting or logging of crime rather than in the crime rate itself. Trend lines illustrating the effect upon rate of industrial accidents of safety training sessions for foremen (Blum, 1949, p. 297) may have been produced by a change in the foreman's reporting standards as a result of the pressure placed upon him to have a good record, rather than by a reduction of actual incidents. *Regression* effects are also usually some orderly function of time and are likewise implausible as explanations of an effect at O_5 greater than the effects at O_2, O_3, and O_4. *Selection* and *mortality* are ruled out in both this design and in Design 2, if the same specific persons are involved at all points. However, if the observations consist of collective products, then a record of the occurrence of absenteeism, quitting, and replacement should be made to make sure that coincidences of personnel change do not provide plausible rival hypotheses to the effects of *X*.

Regarding external validity factors, it is clear that the experimental effect might very well be specific to those populations subject to repeated testing. However, this design is particularly appropriate to those institutional settings in which records are regularly kept and thus constitute a natural part of the environment to which one wants to generalize. Production records in many factories, efficiency reports and other records in the military setting, test performance in classrooms, and the like may represent measurement devices periodically available which would not be reactive, in the sense that they are typical of the universe to which one wants to generalize. On this point, some of Rice's (1931) comments on one of the Industrial Fatigue Research Board experiments introducing rest periods may be in order (Wyatt and Fraser, 1925). Even though they were working in a piece-work shop in which production records were a natural part of the payment procedure, in the actual experiment the interest in precision was such that for a three-week period prior to the experimental change, a group of specially trained observers equipped with stop watches were moved into the plant to record minute-by-minute production rates, to provide more precise data than the gross piece-work pay records would have provided. As Rice (1931) points out, this undoubtedly created a self-consciousness which made the measured behavior less typical of what one wanted to generalize to than it would have been without the additional measurement effort. It seems probable that in this instance the effort to achieve greater precision resulted in a less valid experiment than if the normal production records had been utilized without the workers' awareness of experimentation.

TESTS OF SIGNIFICANCE

If the more advanced sciences use tests of significance less than do psychology and sociology for example, it is undoubtedly because the magnitude and the clarity of effects with which they deal are such as to render tests of significance unnecessary; were our conventional tests of significance to be applied, high degrees of significance would be found. It seems typical of the ecology of the social sciences, however, that they must work the low-grade ore in which tests of significance are necessary. It also seems likely that wherever common sense or intuitive considerations point to a clear-cut effect, some test of significance is possible which formalizes the considerations that have gone into the intuitive judgment. Thus tests of significance must be possible which would distinguish between the several outcomes illustrated in Figure 12.1, judging *A* and *B* to be significant and *E* and *F* not, as far as possible effects of *X* are concerned. Such statistics are not fully developed, but a few approaches are available. Certain conceivable approaches must be rejected as inadequate. If the data in Figure 12.1 represent group means, then a simple significance of the difference between the observations of O_4 and O_5 is insufficient. If in series *E* and *F*, these provided *t* ratios that were highly significant, we would nonetheless not find them interpretable because of the presence of other similar significant shifts occurring on occasions for which we had no matching experimental explanation. Where one is dealing with data such as provided in national opinion surveys, it is common to encounter highly significant shifts from one survey to the next which are nonetheless random noise from the point of view of the interpreting scientist, inasmuch as they represent a part of the variation in the phenomena for which he has no explanation. The effect of a clear-cut event or experimental variable must rise above this level of shift in order to be interpretable. Similarly a test of significance involving the pooled data for all of the pre-*X* and post-*X* observations is inadequate, inasmuch as it would not distinguish between instances of type *E* and instances of type *A*. If one has theoretical expectations appropriate to using a series of points in the postexperimental period (an expectation of an instance like *A* rather than *B* in Figure 12.1), then one might use tests of difference in slope and intercept between the pre-*X* and the post-*X* series (e.g., Walker and Lev, 1953, pp. 390–395), although assumptions of independence of observations are a problem. Mood (1950, pp. 297–298) provides a *t* test, appropriate to effects of type *B* of Figure 12.1, for the significance of the first post-*X* observation from a trend extrapolated from the pre-*X* observations. Usually one would have insufficient data to justify other than a linear fit to the pre-experimental observations, and thus might be misled in an instance such as *G* of Figure 12.1, in which a systematic trend involving a rapid acceleration might be misinterpreted as an effect of the experimental variable except on visual inspection grounds. Since interpretation seems likely to continue to depend in part upon visual inspection, it seems imperative that the plot of the time-series be published. The usual positive autocorrelation for departures from the trend between adjacent points is computable, for a long enough time-series, and perhaps should be taken into account in some significance tests. This autocorrelation provides one reason for ruling out a test of significance for time-series hypothesized to be of type *A* of Figure 12.1, in which one would project the line of best fit from the pre-*X* observations into the post-*X* region, and test the significance of the difference between the predicted and the obtained post-*X* values, using degrees of freedom based upon the number of post-*X* observations.

If such time-series are to be interpreted experimentally, it seems essential that the experimenter specify in advance the expected time relationship between the introduction of the experimental variable and the manifestation of an effect. Were this to be done, the pattern indicated in time-series *D* of Figure 12.1 could be almost as definitive as that in *A*. Exploratory surveys opportunistically deciding upon interpretations of delayed effect would require cross validation before being interpretable. As the time interval between *X* and effect increases, the plausibility of the rival hypotheses of effects from extraneous events increases.

Since this presentation is inevitably exploratory, it seems desirable to look at some actual time-series from which research workers have attempted to infer causation. Figure 12.2*A* presents a long time-series from Farmer (1924), which was interpreted as showing an increase in production rate as a result of a change in work day from ten hours to eight hours in a bottle factory. There is in this time-series a systematic seasonal fluctuation, but the period of observation is long enough to enable us to take this into account to some extent. It perhaps exaggerates the apparent effect that the change was introduced during the summer slump (Farmer reports a correlation of $-.35$ between temperature and production). Note, however, that the summer slump fails to appear in the final summer. Were there more years' data available, one perhaps would do best to remove the seasonal change before proceeding further. This has not been done, however, in the analysis illustrated in Figure 12.2*B*. A model such as *A* or *C* of Figure 12.1 being assumed, separate linear trends have been fitted to the pre-*X* and to the post-*X* portions of the series. These are then compared with a single linear fit for the total series. Even the visual impression of effect is destroyed by the addition of these lines—the total series line obviously fitting the points almost as well as the two separate lines. For a test of significance, the following F-ratio has been computed: The numerator consists of the variance of the points predicted by the two part-series fits from the points predicted by the single linear fit; the denominator, the variance of the observed values from the part-series predictions. The resulting F-ratio is 1.07 with 1 and 27 degrees of freedom, a clearly insignificant result.

This analysis was suggested and carried out by Keith N. Clayton. For those who would like to check the computations or try alternate ones, the numerical values used in the computations were 9.10, 9.58, 8.73, 9.62, 9.20, 9.10, 9.10, 9.53, 10.00, 9.13, 9.62, 10.30, 10.30, 10.38, 9.61, 9.69, 9.52 (experimental change), 9.72, 10.40, 10.98, 10.80, 11.23, 10.90, 10.70, 10.80, 10.13, 11.60, 10.43, 11.32. Mr. Clayton and others have expressed a reservation about the assumption of independence in tests of significance for time-series. To be sure, the time-series of repeated measurements differs from a plot of data in which each point represents a mean of a separate experimental population. One major effect of this difference is to reduce the random variability of points, and to reduce the error term. Since in the method illustrated, the shift is tested against the obtained variability of points, this effect is automatically taken into account. In general, homogeneous nonindependence, orthogonal to experimental treatments, is not the nonindependence to be avoided. The problem of nonindependence arises when it is fully or partially confounded with treatments (when, for example, all plots receiving fertilizer A are adjacent and separate from those receiving fertilizer B). The troubling nonhomogeneous nonindependence in time-series comes from the higher correlation of adjacent points than nonadjacent ones. As mentioned above, this makes unacceptable testing the departure of the whole set of post-X observations from the extrapolated trend line of the pre-X values.

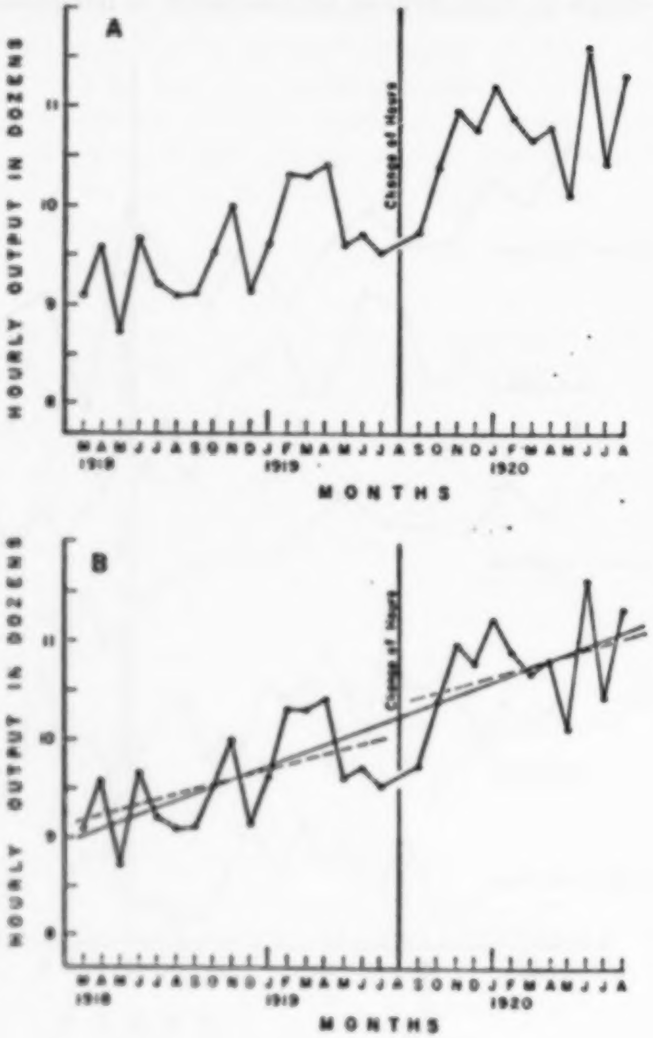


Fig. 12.2.—Change in hourly productivity as a result of shifting from a 10-hour to an 8-hour work day (after Farmer, 1924).

Figure 12.3 takes from Vernon, Bedford, and Wyatt (1924) a set of time-series designed to test the effect of introducing a rest period upon the productivity of various kinds of factory workers. In all except the bottom series, for "Handkerchiefs," the original reporters judged that a positive effect was present. Without intending to discourage the use of time-series experiments, the present writer finds no convincing evidence of an experimental effect in any of them. For "Boots and Shoes," "Bicycle Chains," and "Grooves," there is insufficient pre-X data to establish a trend, particularly considering the manifest instability of production rates. For "Labeling" where the data seems most adequate, the previously illustrated parts-vs.-whole linear fit F-test was applied, resulting in an F of 16.19 with 1 and 35 degrees of freedom, $p < .005$. But visual inspection renders the linear fit implausible, the "effect" seeming to start three observations before the introduction of X, a simple curvilinear or sinusoidal trend being indicated.

There can be noted in all of these series, both before and after X, a general trend toward increased productivity. Perhaps this is just what one should expect of any viable social organization, particularly any organization which is interested in research on the conditions of productivity. Just as an animal always tends, through the many minor adjustments called learning, to a more and more efficient fit to his environment, so a factory is apt to be continually undergoing many minor changes of policy, equipment, and personnel, all having as their goal the correction of things that are not working right, most resulting in some slight improvement. These many sources of improvement are confounded with the *X* under study, as a form of *history*. This general ameliorative trend makes it particularly undesirable to express the effect in terms of mean production rates before and after *X*. It also makes one dubious of interpretations of delayed effects, for the longer the time interval between *X* and supposed effect, the more numerous the extraneous sources of improvement are likely to be.

In preparing our chapter for Gage's *Handbook of Research on Teaching*, Julian Stanley and I considered in detail several possible tests of significance. In the end, equivocation between the alternatives of undesirable weakness and implausible assumptions led us to cut 90 per cent of our prepared discussion from the final manuscript (Campbell and Stanley, 1963). I hope that Professor Stanley will make available his unpublished work along this line, in particular his corrections and extensions of the mode of analysis suggested by the economist Quandt.

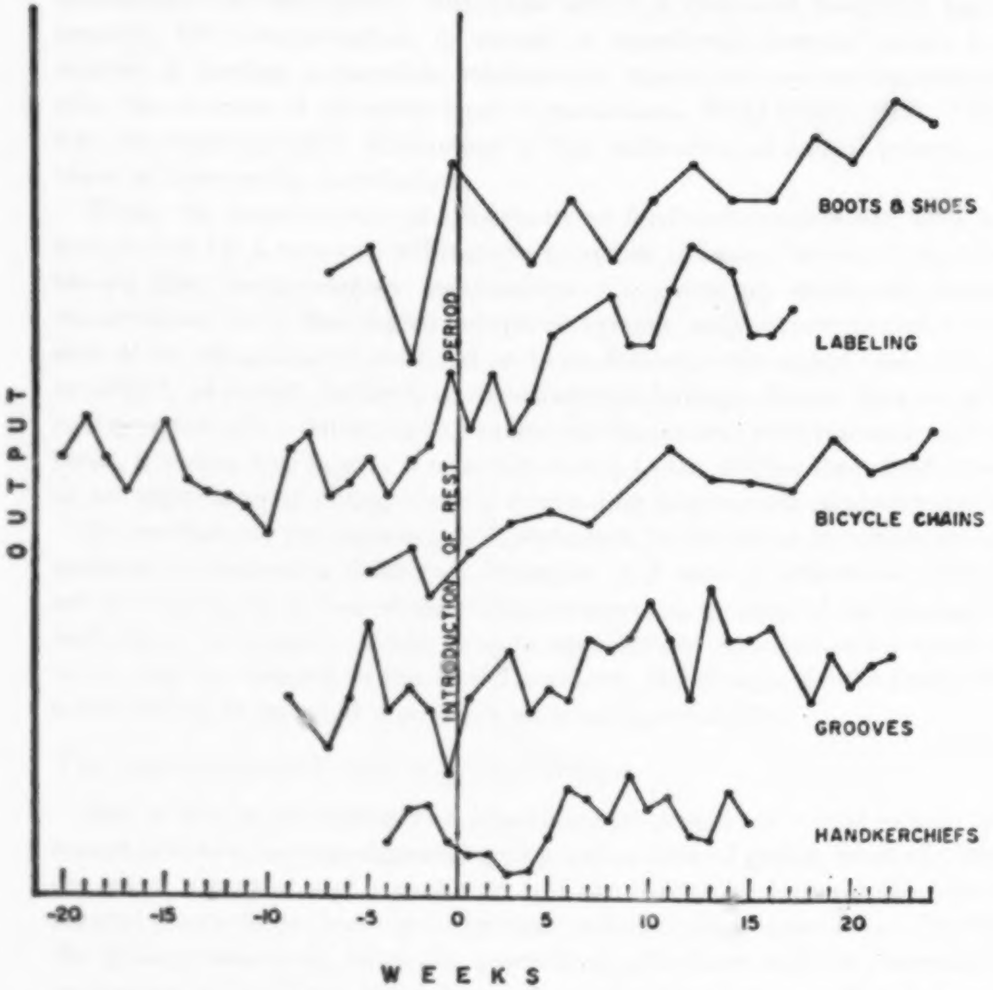


Fig. 12.3.—Effects upon productivity of the introduction of a rest period (after Vernon, Bedford, and Wyatt, 1924).

With regard to linear tests of changes in intercept and/or slope this general defense can be made: The linear test can be regarded as a minimum test. If the data cannot make random fluctuation an implausible rival hypothesis for explaining the effects under these conditions, then the hypothesis that *X* had an effect certainly should be held in abeyance. The linear hypothesis had this useful editorial function in most of the factory production time-series which we examined above. If a body of data passes this test (as did the data on "Labeling"), then comes the stage for ruling out lawlike curvilinear trends as hypotheses explaining the post-*X* gains.

CAN NONEXPERIMENTAL TIME-SERIES BE INTERPRETED AS EXPERIMENTS?

The social scientist, considering the puny nature of the experimental variables which he is likely to be able to manipulate, is ever on the outlook for validly interpretable natural experiments, particularly in situations in which time-series data are regularly available. Cantril (1944) has provided impressive time-series data on public opinion in the United States prior to and during the early years of World War II, in which such catastrophic events as the fall of France seem clearly to have affected public opinion, even though memorable speeches by Roosevelt provide effects of dubious demonstrability. If such interpretive efforts start with a hypothesis about the effect of the natural *X* prior to examination of the results, they are perhaps encompassable within the time-series experiment. But if, as is probably more usual, there is an opportunistic selection of both events to be regarded as *X*'s and portions of the time-series to be interpreted as significantly shifting, one must withhold the precious label of *experiment*. However, if *X* be regarded as a continuously varying quantity, rather than something that occurs in only the two degrees of presence or absence, then the time-series experiment merges with the time-series correlation typical of economic studies. Causal interpretation of such data involves numerous difficulties, including the presence of long-term trends which are associated only by chance, the chance association being repeatedly resampled in the correlated sets of observations with the result of highly significant correlations if significance is based upon the *N* of sampled points. Yule and Kendall (1950) discuss such instances and the particular sampling problem that is involved. Mintz (1946) provides an excellent illustration of the problems involved, in correcting earlier work on the correlation between the rate of lynchings and the price of cotton. In addition, typical problems recur as to the direction of causation and the possibility of correlation without causation. However, through an absence of plausible rival hypotheses, in some cases a temporal correlation is unambiguously interpretable, such as the effect of rainfall upon the price of wheat: Causation is from rainfall to price and not from price to rainfall, or at least was prior to artificial seeding of rain clouds. Such interpretations may be helped by noticing the direction of the temporal lag which maximizes the correlation. But even where a clear-cut temporal lag is present, the interpretation in causal or quasi-experimental terms is a matter of having a plausible explanatory model for one interpretation, plus the absence of plausible rival explanations. Wold (1953, 1954, 1956) has provided valuable discussions of the utilization of causal interpretations of time-series correlations.

While the introduction of cybernetic or feedback models has been accompanied by a renewed willingness to speak in causal terms, it has also shown how compensatory mechanisms can cover up important causal connections. In a thoroughly adaptive system, only experimental intrusion of an unnatural or artificial sort can delineate the causal connections involved, as recent research in experimental biology shows. And on general grounds of avoiding capitalization on chance and post-hoc rationalizations, it seems well to give a separate status to the deliberate introduction of an experimental change into a continuing time-series of observations.

In conclusion, the time-series experiment, in the form in which an experimenter controls a deliberate intrusion of X into an otherwise uniform set of conditions, is one of the frequent research designs of the biological and physical sciences. While to date appropriate and successful applications may be lacking in the social sciences, the design should probably, nonetheless, be granted a position of semirespectability.

The nonequivalent control-group design

One of the most widespread experimental designs in social science research involves an experimental group and a control group, both of which are given a pretest and a post-test, but the control group and the experimental group do not have pre-experimental sampling equivalence. Rather, the groups constitute naturally assembled collectives such as classrooms, as similar as available but yet not so similar that one can dispense with the pretest. The assignment of X to one group or the other is ideally random and under the experimenter's control.

$$\begin{array}{ccc} O & X & O \\ \hdashline O & & O \end{array}$$ Nonequivalent control-group design.

Two things need to be kept clear about this design: First, it is not to be confused with a design in which experimental subjects are assigned *randomly* from a common population to the experimental and the control groups. Second, in spite of this, Design 10 should be recognized as well worth using in many instances in which true experimental designs are impossible. In particular it should be recognized that the addition of even an unmatched or nonequivalent control group reduces greatly the equivocality of interpretation over what is obtained in Design 2, the one-group pretest–post-test design. The more similar the experimental and the control groups are in their recruitment, and the more this similarity is confirmed by the scores on the pretest, the more effective this control becomes. When these desiderata are approximated, for purposes of internal validity the design can be regarded as controlling the main effects of *history*, *maturation*, *testing*, and *instrumentation*, in that the difference for the experimental group between pretest and post-test (if greater than that for the control group) cannot be explained by main effects of these variables such as would affect both the experimental and the control groups.

An effort to explain away a gain specific to the experimental group in terms of such extraneous factors as *history*, *maturation*, or *testing* must hypothesize an interaction between these variables and the specific selection differences that distinguish the experimental and control groups. While in general such interactions are unlikely, there are a number of situations in which they might be invoked. Perhaps most common are interactions involving *maturation*. If the experimental group consists of psychotherapy patients and the control group some other handy population tested and retested, a gain specific to the experimental group might well be interpreted as a spontaneous remission process specific to such an extreme group, a gain that would have occurred even without X. Such a *selection-maturation interaction* (or a *selection-history interaction*, or a *selection-testing interaction*) could be mistaken for the effect of X, and thus represents a threat to the *internal* validity of the experiment. This possibility has been represented in the eighth column of Table 12.1.

The illustration of psychotherapy applicants provides an instance in which the assumptions of homogeneous regression and of sampling from the same universe except for extremity of scores would seem likely to be inappropriate. The inclusion of normal controls in psychotherapy research is of some use, but extreme caution must be employed in interpreting results. It seems important to distinguish two versions of Design 10, and to give them different status as approximations to true experimentation. On the one hand, there is the situation in which the experimenter has two natural groups available, *e.g.*, two classrooms, and has free choice in deciding which gets *X* or at least has no reason to suspect differential recruitment related to *X*. Even though the groups may differ in initial means on *O*, the study may approach true experimentation. On the other hand, there are instances of Design 10 in which the respondents clearly are self-selected, the experimental group having deliberately sought out exposure to *X*, with no control group available from this same population of seekers. In this latter case, the assumption of uniform regression between experimental and control groups becomes less likely, and *selection-maturation interaction* (and other selection interactions) becomes more probable. The "self-selected" Design 10 is thus much weaker. Nonetheless, it may provide information which in many instances would rule out the hypothesis that *X* has an effect. The control group, even if widely divergent in recruitment and mean level, assists in the interpretation. The equivocality produced by the assumption of homogeneity of regression becomes a particular problem if Design 10 produces seeming evidence of significant effect.

The scale used in quantifying measures can in some circumstances provide an inhomogeneity of regression misinterpretable as an effect of *X* in Design 10. Consider a case in which the control group and the experimental group differ widely in level, with the control group showing a slight gain, the experimental group showing a large gain. In such a case, a logarithmic transformation of the metric justifiable on other grounds might entirely remove the difference in slopes. In such a case, the character of the initial metric, combined with the difference in the initial levels of the groups, explains the effect. To assimilate this possibility into the check-off list, metric can be included as one of the aspects of *instrumentation*, this threat thus becoming a *selection-instrumentation* interaction. I am indebted to Dr. Leroy Wolins for this observation.

The multiple time-series design

While No. 10, the nonequivalent control-group design, is an experimental design worthy of more detailed discussion because of its ubiquitous availability, its discussion here has been introduced primarily as a transition to the multiple time-series design. Let us illustrate a transition from Design 2 through Design 10 to Design 14 with the content of E. L. Thorndike's classic quasi-experiments on formal discipline. Let us assume that Thorndike found a course in Latin associated with a significant gain in English vocabulary, as modern re-analysis with covariance might well show, for note the trends in Broyler, Thorndike, and Woodyard (1927). A pre-Thorndikian study might be graphed as in Figure 12.4*A*. Thorndike adds a control group, and in overcoming the pretest differential, at least partially avoids the matching fallacy by testing post-test scores against regressed pretest scores. The hypothetical results can be graphed as in Figure 12.4*B*. Adding this control removes Design 2's plausible rival hypotheses (sources of invalidity) of the main effects of *history*, *maturation*, and *testing*. Let us suppose that the sophomores taking Latin do gain in vocabulary over the course of the year significantly more than do the non-Latin-taking "controls." Is this attributable to the effect of

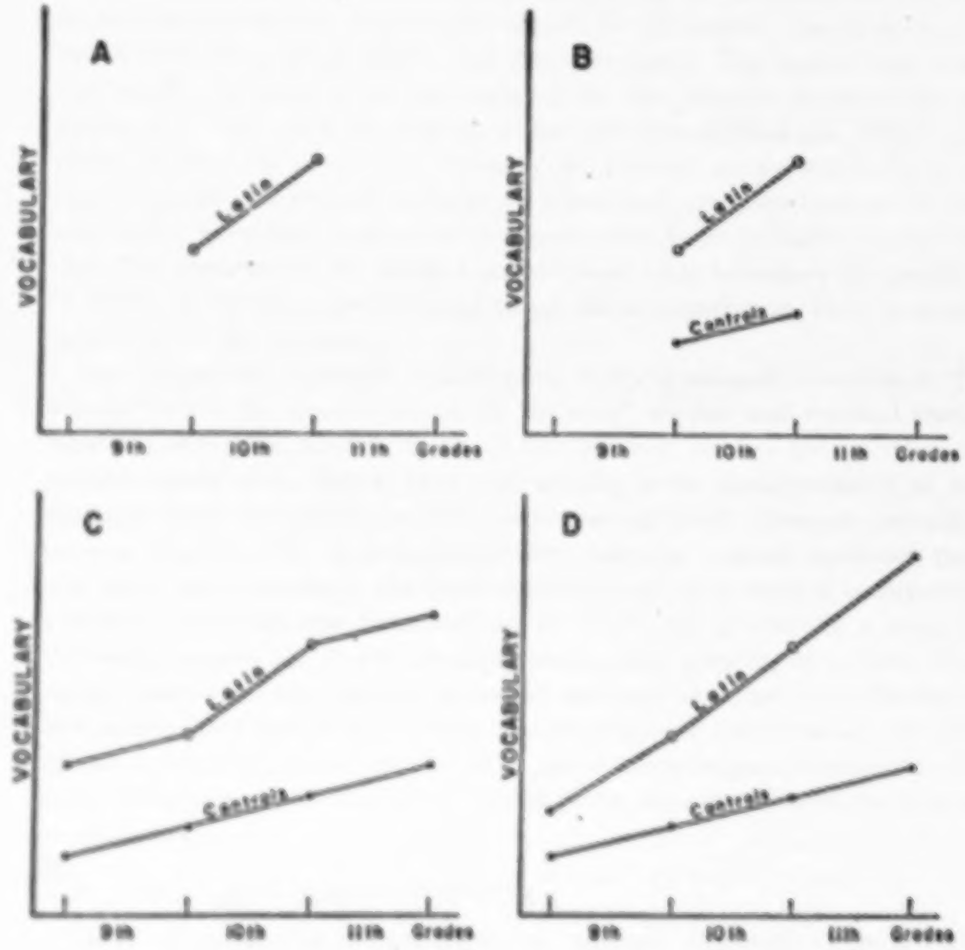


Fig. 12.4.—Forms of quasi-experimental analysis for the effect of specific course work, including multiple time-series.

Latin without plausible rival? Of course not, for it is very plausible that the students selecting Latin were not only above average in English vocabulary to begin with, but also differed from the others in usual annual rate of increase: i.e., that they would have shown greater gains than the controls even without Latin. Hence the minus in the eighth column of Table 12.1 for Design 10. But this assumption of differential slope or differential regression has implications which are themselves testable. And it is the very essence of the quasi-experimental program that all the data available be brought to bear on the problem of rendering implausible the uncontrolled factors, with, if necessary, "patched-up" designs of great inelegance resulting (viz., Campbell and Stanley, 1963, Design 15). In the present instance a very useful alternative is available, i.e., Design 14, the multiple time-series. Presumably we may have vocabulary tests for several previous years. If these data look like Figure 12.4*D*, then we decide that Latin has had no effect. But if the outcome is as in Figure 12.4*C*, we may conclude that Latin (here assumed to be a one-year course only) has improved English vocabulary. This is not as good as an experiment using randomization in assigning class schedules. Nonetheless, we have in Table 12.1 a solid row of pluses for internal validity. Each one of those eight rival hypotheses has been rendered implausible. The design is very distinctly better than Design 10, and is, in fact, one of the very best of quasi-experimental designs.

Technically speaking, the column-eight plus should not be completely solid. Interactions of *selection* and *history*, or of *selection* and *maturation* and *history*, are still possible rivals, and one would want to scrutinize the total environments to judge the plausibility of these. Such possible rivals would exist if those choosing Latin were transferred to a different school on that account, or even if Latin were a freshman course so that change of school for both groups coincided with Latin, or if the decision not to take Latin involved a formal renunciation of a college career, so that the vocabulary-learning motivation of the "controls" underwent a specifiable change, etc. But with a properly selected control group and setting, these interactions might very well end up being judged implausible.

The multiple time-series design can also be reached through a successive improvement of Design 2, the one-group pretest-post-test design, via Design 7, the interrupted time-series design. In this sequence, the addition of the control series to Design 7 removes the minus from the *history* column.

Note that the control series need not be separate groups of persons. Control series differing in contents sampled may be useful. Thus in the advertising evaluation of a novel and intense sales campaign, the sales of rival brands could provide the control series. Or in our case, the English vocabulary can be split into groups by root. If the plots for a "control group" of Anglo-Saxon roots showed no gains, while those of Latin did, the rival hypothesis of differential rates of autochthonous vocabulary growth would not be plausible. Thorndike (Thorndike and Ruger, 1923) used such a comparison, although in a single pretest-type design and for a different purpose.

At the present time tests of significance appropriate to this design have not been uncovered. The issues of autocorrelation or greater shared error for adjacent points (at least within experimental group or within control) are still present. A new feature, adding precision if usable, is the cross correlation of zero lag between experimental group and control. To estimate this correlation, it probably should be computed separately for the pre-X and the post-X series, and then averaged. The higher this value, the smaller the error term, presumably, for the jump in the experimental group at X. But such an error term has not been spelled out. Where it is plausible that the same secular and cyclical trends are present in both the experimental and control series (e.g., when both are simultaneous in time and under presumed control of the same exogenous variables except X), then the presence of the control series would seem to reduce the problem of selecting the appropriate trend to use for extrapolation. How to implement this is the problem.

One suggested approach is analogous to the economist's confident "removal" from the time-series of all "known" secular and cyclical trends before looking for novel effects. While in their studies this is done by extrapolating to a degree that will usually seem undependable in our settings (and that often involves regression artifacts, through providing an expectation that systematically lies closer to central tendency than will obtained extremes), the control-group data may serve a comparable function. Thus we can "remove" these trends by generating a series of difference scores (or if the absolute levels vary greatly, of ratios). This single time-series can then be tested by methods appropriate to Design 7, the single interrupted time-series. Assumptions of linear trend (or even of linear trend plus zero slope) will be much more frequently plausible for such difference scores than they would be for the raw data of the experimental series.

The cross-lagged panel correlation

In the discussion of the possibility of "natural" time-series experiments above, the use of lagged cross-series correlations was suggested as one means of deciding the direction of causation. Where two data series corriate, e.g., annual Norwegian harvest yields and numbers of marriages (Yule and Kendall, 1950), the direction of causation may be equivocal (given the universally hypothesized efficacy of fertility rites in increasing harvest yields, etc.). In such a situation, $r_{C_oE_{o+1}}$ should be greater than $r_{C_{o+1}E_o}$, where *C* stands for cause, *E* for effect. These cross-lagged series correlations can frequently differentiate the relative plausibilities of competing causal interpretations. When both variables are on both sides of the comparison, i.e., when relative correlation magnitude is used rather than the absolute level of $r_{C_oE_o}$, secular trends or long-term cycles are controlled.

The correlations involved in these analyses are correlations between two variables over a population of temporally distributed occasions. They are correlations of Cattell's type *P*, with lags (see the papers of Cattell and Holtzman in this volume). Rarely will we have data sufficient to justify these. Typically our situation is one of annual measurements repeated over a very few years. But typically also, we have data for each year on many individuals. In such a situation, the comparison of cross-lagged correlations computed over a population of persons seems useful. For the economist, these would be lagged cross-sectional correlations. In the harvest-marriage example, these would be correlations of marriage rates and harvest magnitudes across a population of nations. The cross-lagging would involve correlating harvest rates for one year with marriage rates for the next year, and vice versa, with an *N* of nations. Our criterion becomes $r_{C_1E_2} > r_{C_2E_1}$.

We have now moved into an area where social science data are more readily available. Longitudinal studies of child development, archives of school records, panel studies of voting and of consumer preferences, all provide appropriate settings, with an *N* of individuals, parent-child pairs, etc. Do adjustment problems cause poor school performance or vice versa? Does lack of parental love cause children to be behavior problems, or does a difficult child cause parents to love less? While in many such instances, the causal relations are doubtless in both directions, an index of relative preponderance would be very valuable, and where a preponderance is clear, the status of the dominant hypothesis is clearly enhanced, and the credibility of the weaker one must be based upon other bodies of data.

The cross-lagged panel correlation when applied to dichotomous data corresponds to Lazarsfeld's "Sixteenfold Table" (Lazarsfeld, 1947; Levenson, 1955; Lipset, Lazarsfeld, Barton, and Linz, 1954). Conversely, the cross-lagged panel correlation would seem to be the most feasible means of extending the "Sixteenfold Table" beyond the dichotomous situation.

Although Lazarsfeld has not seen fit to present this method publicly in spite of having an eminently publishable 24-page manuscript, I judge this to be the most usefully ingenious of his many methodological contributions. In selecting the name, I have employed the term "panel" to indicate the intellectual heritage of the analysis, as well as to distinguish it from the cross-lagged series correlation.

Let us illustrate the logic of Lazarsfeld's design with some hypothetical data typical of the setting in which Lazarsfeld first introduced it. In a political campaign, Own Vote Intention and Expectation as to the Winner are correlated. Is this due to wishful thinking, with Vote Intention causing Expectation? Or is it due to the bandwagon effect, with Expectation causing Vote Intention? Let us suppose that in two waves of a panel study we find this correlation increasing, as in Table 12.2. The equivocality of ordinary correlational data, and the ingenuity of Lazarsfeld's analysis, becomes apparent if we note that among the many shifts which would have made the transformation possible are the polar opposites shown in Table 12.3. Here we have presented only 8 of the potential 16 cells. We have neglected the other 8 which reduce the correlation. Lazarsfeld's type of presentation, for the bandwagon effect outcome, is shown in Table 12.4. The empty parentheses indicate the potential loci of the "wishful thinking" instances, nonexistent in this pure case. While it is obvious that we are observing a "causal" antecedence-subsequence relation only for those 20 cases that change, we probably have no grounds for expecting them to be different from those who on the first interview

TABLE 12.2

Correlation Increase between Expected Winner and Vote Intention with Hypothetical Panel Data

July Survey

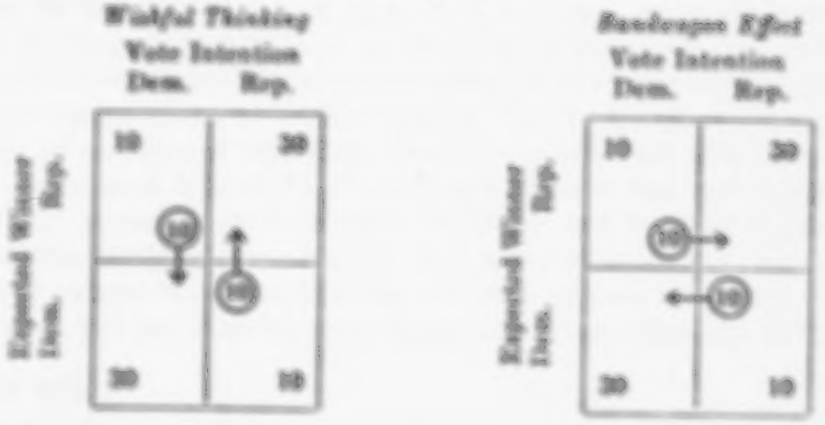
Expected Winner	Vote Intention Dem.	Vote Intention Rep.
Rep.	20	30
Dem.	30	20

September Survey

Expected Winner	Vote Intention Dem.	Vote Intention Rep.
Rep.	10	40
Dem.	40	10

TABLE 12.3

Two Modes of Shift for the Effect in Table 12.2



Wishful Thinking
Vote Intention
Dem.
Rep.
Expected Winner
Rep.
Dem.
10
30
10
10
30
10
Bandwagon Effect
Vote Intention
Dem.
Rep.
Expected Winner
Rep.
Dem.
10
30
10
10
30
10

already occupied the comfortable Dem.: Dem. and Rep.: Rep. diagonal corners where expectation and vote are in agreement.

Diagramed as cross-lagged panel correlations, the bandwagon outcome provides the two fourfold tables shown in Table 12.5.

Lazarsfeld's "index of mutual effects" is a complex combination of ratios of the determinants of changes to the base of changes, separating changes increasing the correlation from changes decreasing the correlation. Since I am unfamiliar with the mathematical techniques employed, and since these are unlikely to be usable for nondichotomous variables, I shall not duplicate them here, but will illustrate the present suggested analysis for the three bodies of data which Lazarsfeld presents in full. The first asks a question of the 1940 voting data (otherwise presented in Lazarsfeld, Berelson, and Gaudet, 1944): Did personal attraction to Willkie cause people to vote Republican, or did Republicanism rather cause people to like Willkie. The basic data are as shown in Table 12.6.

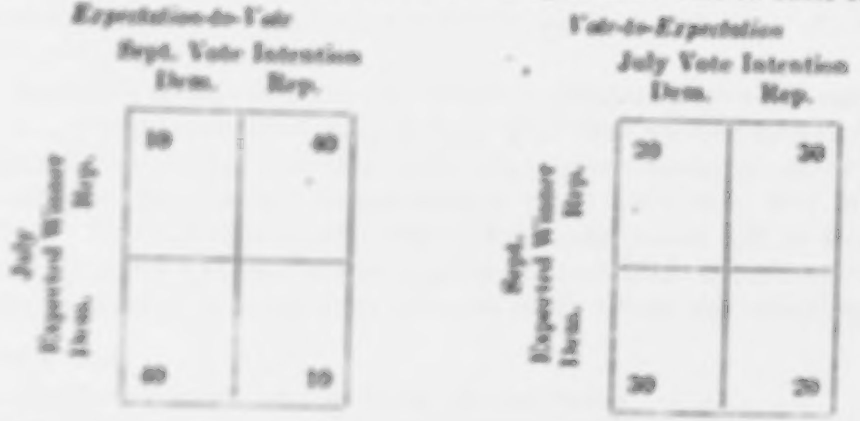
TABLE 12.4

Sixteenfold Table Presentation of Bandwagon-Effect Case of Table 12.3

July interviews	September interviews			
Vote : Expect.	Rep. : Rep.	Rep. : Dem.	Dem. : Rep.	Dem. : Dem.
Rep. : Rep.	30	0	0	0
Rep. : Dem.	10	10	0	10
Dem. : Rep.	10	0	10	10
Dem. : Dem.	0	0	0	30

TABLE 12.5

Cross-Lagged Panel Correlations for the Bandwagon-Effect Case of Table 12.3



Expectation-to-Vote
Sept. Vote Intention
Dem.
Rep.
July
Expected Winner
Rep.
Dem.
10
40
40
10
Vote-to-Expectation
July Vote Intention
Dem.
Rep.
Sept.
Expected Winner
Rep.
Dem.
20
30
30
20

For these data, the correlation between vote intention and liking of Willkie moves from .53 on the first interview to .67 on the second. The cross-lagged panel correlations are $r_{V_1L_2} = .65$, whereas $r_{L_1V_2} = .51$. These differ significantly at the .015 level (two-tails, z transformation, intercorrelation of correlation arrays disregarded, $N = 266$). Lazarsfeld's index for this instance is .129. This index value, but not the mode of computing it, is presented in the chapter on voting in the *Handbook of Social Psychology* (Lipset, Lazarsfeld, Barton, and Linz, 1954). It is concluded that Republicanism causes more liking of Willkie than vice versa.

Earlier (Campbell and Clayton, 1961) I had argued that the differential in a cross-lagged panel correlation could be explained away in some instances by two concomitant plausible rival hypotheses, to wit: (*a*) That increased correlation ($r_{V_1L_1} < r_{V_2L_2}$) was required and could be due to extraneous sources such as the test-retest effect and other pressures for consistency; (*b*) that the asymmetry could be due to the higher reliability

TABLE 12.6

Lazarsfeld's Sixteenfold Table for Vote Intention and Attitude toward Willkie

First interview	Second interview			
Vote : Attitude	Dem. : Ag.	Dem. : For	Rep. : Ag.	Rep. : For
Dem. : Against Willkie	68	2	1	1
Dem. : For Willkie	11	12	0	1
Rep. : Against Willkie	1	0	23	11
Rep. : For Willkie	2	1	3	129

or stability of one of the variables. In cases where $r_{C_1E_2} > r_{C_2E_1}$ it may usually be found that $r_{C_1C_2} > r_{E_1E_2}$ if $r_{C_1E_1} < r_{C_2E_2}$ as in the pure case above. In the present data, $r_{V_1V_2} = .94$, while $r_{W_1W_2}$ is only .77. I no longer believe this lower reliability provides a plausible rival hypothesis however. The correlation shows test-retest stability for the period of change, and if taken as an index of reliability, should be generalized alike to both V_1 and V_2, or to both W_1 and W_2. Even though we know that various attitude topics vary greatly in their intrinsic reliability, and may find it plausible that Willkie attitude is less reliable than party preference, we would still normally assume that the reliability of Willkie attitude was equal on both interviews, and the same for party preference. Thus reliability differences between the variables, as reliability is normally conceived of, enter in to lower both $r_{V_1W_2}$ and $r_{V_2W_1}$ equally, and thus do not provide an explanation of one's being larger than the other. However, a shift in reliability of one of the variables (ascertained by other means than test-retest) may be in some instances a plausible rival hypothesis. A variable which increases in reliability from Time 1 to Time 2 will, *ceteris paribus*, show up as an "effect" rather than a "cause."

TABLE 12.7

Lazarsfeld's Sixteenfold Table on Vote Intention and Expectation

First interview	Second interview			
Vote:Expect.	Rep.:Rep.	Rep.:Dem.	Dem.:Rep.	Dem.:Dem.
Rep.:Rep.	59	10	1	0
Rep.:Dem.	13	14	0	4
Dem.:Rep.	3	0	1	3
Dem.:Dem.	6	3	7	69

A second body of data presented by Lazarsfeld deals with the Bandwagon vs. Wishful Thinking hypothesis, and is shown in Table 12.7. In this case, the correlation between Expectation and Vote Intention remains nearly constant: $r_{E_1V_1} = .63$, $r_{E_2V_2} = .64$. However, in agreement with Lazarsfeld's analysis $r_{E_1V_2} = .62$ which is larger than $r_{E_2V_1} = .54$. The difference between correlations is significant only at the $p < .24$ level. Lazarsfeld's index is .084, and while he says "Thus we see that the 'bandwagon effect' predominates" (Lazarsfeld, 1947, p. 15), it is unlikely that the index would have been significant by the test provided by William S. Robinson in Appendix C of that paper. For the benefit of interest generated in this specific content problem, it should be noted that in the subsequent voting panel study (Berelson, Lazarsfeld, and McPhee, 1954, p. 289) they report that neither that study nor the one we have just considered showed a differential influence on the bandwagon effect vs. wishful thinking analysis.

These data do indicate, however, that it is possible to meet the criterion of $r_{C_1E_2} > r_{C_2E_1}$ even when $r_{C_1E_1} \geq r_{C_2E_2}$. How shall we interpret such instances? Note in this case that unlike the model initially presented, the stability of the "cause," Expectation, is lower (.60) than that of the "effect," Vote Intention (.82). The causal model would still be tenable for an intermittent unstable cause, and in instances like rainfall and wheat price, or harvest and marriage rate, we could accept the causal model even where the general level of correlation did not increase year after year, and even where cause was less reliable than effect.

TABLE 12.8

Lazarsfeld's Sixteenfold Table on Seeing Ads and Buying

First interview	Second interview			
See:Buy	See:Buy	See:Don't Buy	Don't See: Buy	Don't See: Don't Buy
See:Buy	83	8	35	7
See:Don't Buy	22	68	11	28
Don't See:Buy	25	10	95	15
Don't See:Don't Buy	8	32	6	493

In a third case, Lazarsfeld examines the equivocal correlation between buying a product and seeing the ads for the product. As he points out, and as Festinger (1957) has more recently emphasized, it could well be that owning or using a product increases one's attentiveness to its advertisements. Lazarsfeld's data come from a two-wave panel study of advertising exposure and commodity purchase. These data are shown in Table 12.8. For these data, $r_{S_1B_1} = .29$, $r_{S_2B_2} = .32$. The cross-lagged panel correlations are $r_{S_1B_2} = .37$, $r_{S_2B_1} = .27$, this difference being significant at the .016 level. The reliabilities are $r_{S_1S_2} = .59$, $r_{B_1B_2} = .78$. Here again the "cause" is less stable than the "effect." The Lazarsfeld index is in this instance .077.

For this design in its dichotomous form, regression probably becomes a plausible rival hypothesis when the item marginals are extreme and differ for the two variables, both for the Lazarsfeld index and for the phi coefficients of the lagged cross correlation as they are differentially affected by the ceilings created by the uneven marginals. However, for continuous variables, no such limitation seems likely. The bugaboo to all correlational interpretation represented by third variable causation could possibly, in complex form, provide plausible rival hypotheses in some instances, and in any application these should be examined. A shift in reliability of one of the variables would lead to mistaken inference if it went unnoticed. In general, however, the design represents a welcome addition to our armamentarium of quasi-experimental analyses. Using it on the bodies of data available and exposing the interesting results to the criticisms of colleagues wise in the specific content areas will in the long run enable us better to fill out the check-list of plausible rival hypotheses.

Summary

The problem of moving from description of change to inferring the causes of change has been examined in terms of the threats to internal validity found in quasi-experimental designs. While change is most interpretable where the experimenter has intruded into natural processes, there designs available to the unintrusive observer have been found worthy of application and careful examination. These three are the interrupted time-series design, the multiple time-series design, and the cross-lagged panel correlation. In all such assemblies of nonexperimental data, quasi-experimental interpretation depends upon rendering implausible those explanations of the observed change which stand as rivals to the hypothesis that a designated experimental variable caused the change. A check-list of eight classes of frequently plausible rival hypotheses has been used in the examination of the data arrangements under discussion.

References

Index

PRECEDING PAGE BLANK-NOT FILMED

References

Acton, F. S. 1959. *Analysis of straight-line data.* New York: Wiley.

Alexander, H. W. 1946. A general test for trend. *Psychol. Bull.*, **43**, 533–555.

Allport, G. W. 1937. *Personality: A psychological interpretation.* New York: Holt.

Anastasi, A. 1958. *Differential psychology* (3rd ed.). New York: Macmillan.

Anderson, N. H. 1961. Scales and statistics: Parametric and nonparametric. *Psychol. Bull.*, **58**, 305–316.

Anderson, R. L. 1942. Distribution of the serial correlation coefficient. *Ann. math. Statist.*, **13**, 1–13.

Anderson, R. L., and Bancroft, T. A. 1952. *Statistical theory in research.* New York: McGraw-Hill.

Anderson, T. W. 1958. *Introduction to multivariate statistical analysis.* New York: Wiley.

———, 1960. Some stochastic process models for intelligence test scores. In Arrow, K. J., Karlin, S., and Suppes, P. (Eds.), *Mathematical methods in the social sciences.* Stanford: Stanford University Press.

———, 1961. *The use of factor analysis in the statistical analysis of multiple time series.* Tech. Rep. No. 12, Contract AF 41 (657)-214, Sch. Aviat. Med., Brooks A. F. Base.

———, 1962. The choice of the degree of polynomial regression as a multiple decision problem. *Ann. math. Statist.*, **33**, 255–265.

Bargmann, R. 1957. *A study of independence and dependence in multivariate normal analysis.* Chapel Hill: University of North Carolina, Inst. Statist., Mimeo. Ser.

Bartlett, M. S. 1935. Some aspects of the time-correlation problem in regard to tests of significance. *J. Roy. Statist. Soc.*, **98**, 536–543.

———, 1951. The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Ann. Eugen.*, **16**, 199–214.

Beck, S. J. 1953. The science of personality: Nomothetic or idiographic? *Psychol. Rev.*, **60**, 353–359.

Bereiter, C., and Freedman, M. B. 1962. Fields of study and the people in them. In Sanford, N. (Ed.), *The American college.* New York: Wiley.

Berelson, B. R., Lazarsfeld, P. F., and McPhee, W. N. 1954. *Voting.* Chicago: University of Chicago Press.

Blum, M. L. 1949. *Industrial psychology and its social foundations.* New York: Harper.

Bock, R. D. 1960. Components of variance analysis as a structural and discriminal analysis for psychological tests. *Brit. J. statist. Psychol.*, **13**, 151.

———. 1961. *Fixed-point subroutines for multivariate statistical computation.* Chapel Hill: University of North Carolina, Psychometric Lab., Res. Memo. No. 6.

———. 1962. *Programming univariate and multivariate analysis of variance.* Chapel Hill: University of North Carolina, Psychometric Lab., Res. Memo. No. 11.

Boneau, C. A. 1960. The effect of violations of assumptions underlying the *t* test. *Psychol. Bull.*, **57**, 49–65.

Box, G. E. P. 1950. Problems in the analysis of growth and wear curves. *Biometrics*, **6**, 362–389.

———. 1952. Multifactor designs of the first order. *Biometrika*, **39**, 49–57.

———. 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. math. Statist.*, **25**, 484–498.

Box, G. E. P., and Hay, W. A. 1953. A statistical design for the efficient removal of trends occurring in a comparative experiment with an application in biological assay. *Biometrics*, **9**, 304–319.

Brownlee, K. A. 1960. *Statistical theory and methodology in science and engineering.* New York: Wiley.

Brolyer, C. R., Thorndike, E. L., and Woodyard, E. 1927. A second study of mental discipline in high school studies. *J. educ. Psychol.*, **18**, 377–404.

Campbell, D. T. 1957. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.*, **54**, 297–312.

Campbell, D. T., and Clayton, K. N. 1961. Avoiding regression effects in panel studies of communication impact. In *Studies in public communication*, No. 3. Chicago: University of Chicago, Dept. of Sociol.

Campbell, D. T., and Stanley, J. S. 1963. Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

Cantril, H. 1944. *Gauging public opinion.* Princeton: Princeton University Press.

Cattell, R. B. 1946. *The description and measurement of personality.* New York: World.

———. 1951*a*. On the disuse and misuse of *P*, *Q*, *Q′* and *O* techniques in clinical psychology. *J. clin. Psychol.*, **7**, 203–214.

———. 1951*b*. *P-technique, a new method for analyzing the structure of personal motivation. Trans. N. Y. Acad. Sci.*, Ser. II, **14**, 29–34.

———. 1952*a*. *Factor analysis.* New York: Harper.

———. 1952*b*. The three basic factor-analytic research designs—their interrelations and derivatives. *Psychol. Bull.*, **49**, 499–520.

———. 1953. A quantitative analysis of the changes in the culture pattern of Great Britain, 1837–1937, by *P*-technique. *Acta Psychol.*, **9**, 99–121.

———. 1955. The chief invariant psychological and psychophysical functional unities found by *P*-technique. *J. clin. Psychol.*, **11**, 319–343.

———. 1957. *Personality and motivation structure and measurement.* New York: Harcourt, Brace, and World.

———. 1961*a*. Group theory, personality and role: A model for experimental researches. In Geldard, R. (Ed.), *Defence psychology.* New York: Pergamon Press.

———. 1961*b*. The theory of situational, instrument, second order, and refraction factors in personality structure research. *Psychol. Bull.*, **58**, 160–174.

———. 1962. *The theory of fluid and crystallized intelligence: A crucial experiment.* Urbana: University of Illinois, Lab. Pers. Assess. and Group Behav.

———. (Ed.). 1963. *Handbook of multivariate experimental psychology.* Chicago: Rand McNally.

Cattell, R. B., and Adelson, M. 1951. The dimensions of social change in the U.S.A. as determined by *P*-technique. *Soc. Forces*, **30**, 190–201.

Cattell, R. B., Cattell, A. K. S., and Rhymer, R. M. 1947. *P*-technique demonstrated in determining psycho-physiological source traits in a normal individual. *Psychometrika*, **12**, 267–288.

Cattell, R. B., and Cross, K. 1952. Comparison of the ergic and self-sentiment structures found in dynamic traits by *R*- and *P*-techniques. *J. Pers.*, **21**, 250–271.

Cattell, R. B., and Luborsky, L. B. 1950. *P*-technique demonstrated as a new clinical method for determining personality and symptom structure. *J. gen. Psychol.*, **42**, 3–24.

Cattell, R. B., and Nesselroade, J. 1962. *The IPAT seven factor state battery.* 1602 Coronado Dr., Champaign, Ill.: IPAT.

Cattell, R. B., and Scheier, I. H. 1961. *The meaning and measurement of neuroticism and anxiety.* New York: Ronald Press.

Cattell, R. B., and White, O. 1962. *Confactor rotation: Some unsolved problems.* Urbana: University of Illinois, Lab. Pers. Assess. Group Behav.

Cattell, R. B., and Williams, H. F. 1953. *P*-technique, a new statistical device for analyzing functional unities in the intact organism. *Brit. J. prev. & soc. Med.*, **7**, 141–153.

Chapin, F. S. 1955. *Experimental designs in sociological research* (Rev. ed.). New York: Harper.

Collier, R. M. 1944. The effect of propaganda upon attitude following a critical examination of the propaganda itself. *J. soc. Psychol.*, **20**, 3–17.

Cornfield, J., and Tukey, J. W. 1956. Average values of mean squares in factorials. *Ann. math. Statist.*, **27**, 907–949.

Courtis, S. A. 1950. *Maturation units and how to use them.* 9110 Dwight Ave., Detroit, Michigan: Author.

Cox, D. R. 1957. The use of a concomitant variable in selecting an experimental design. *Biometrika*, **44**, 150–158.

———. 1958. *Planning of experiments.* New York: Wiley

Crook, M. N. 1937. The constancy of neuroticism scores and self-judgments of constancy. *J. Psychol.*, **4**, 27–34.

Danford, M. B., and Hughes, H. 1957. Mixed model analysis of variance assuming equal variances and equal covariances. *USAF Sch. Aviat. Med. Rep.*, No. 57-144.

Danford, M. B., Hughes, H., and McNee, R. C. 1960. The analysis of repeated measurements experiments. *Biometrics*, **16**, 547–565.

Davis, F. B. 1961. The assessment of change. In Marquette University, Reading Center, *Tenth yearbook, national reading conference*. Milwaukee: Marquette University.

DuBois, P. H., and Manning, W. H. 1961. *Methods of research in technical training.* St. Louis: Washington University, Tech. Rep. No. 3, ONR No. Nonr 816(02).

Edwards, A. L. 1960. *Experimental design in psychological research* (Rev. ed.). New York: Holt, Rinehart, and Winston.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. *Biometrics*, **3**, 1–21.

Faddeeva, V. N. 1959. *Computational methods of linear algebra.* New York: Dover.

Farmer, E. 1924. *A comparison of different shift systems in the glass trade.* Rep. No. 24, Med. Res. Council, Indus. Fatigue Res. Bd. London: His Majesty's Stationery Office.

Feldt, L. S. 1958. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, **23**, 335–353.

Ferguson, G. A. 1959. *Statistical analysis in psychology and education.* New York: McGraw-Hill.

Festinger, L. 1957. *A theory of cognitive dissonance.* Evanston: Row Peterson.

Fisher, R. A. 1951. *The design of experiments* (6th ed.). New York: Hafner.

Fisher, R. A., and Yates, F. 1957. *Statistical tables for biological, agricultural, and medical research.* New York: Hafner.

Fleishman, E. A., and Hempel, W. E. 1954. Changes in factor structure of a complex psychomotor task as a function of practice. *Psychometrika*, **19**, 239–252.

Gaito, J. 1958a. The Bolles-Messick coefficient of utility. *Psychol. Rep.*, **4**, 595–598.

———, 1958b. The single Latin square in psychological research. *Psychometrika*, **23**, 369–378.

———, 1958c. Statistical dangers involved in counterbalancing. *Psychol. Rep.*, **4**, 463–468.

———, 1959. Non-parametric methods in psychological research. *Psychol. Rep.*, **5**, 115–125.

———, 1960. Expected mean squares in analysis of variance techniques. *Psychol. Rep.*, **7**, 3–10.

———, 1961. Repeated measurements designs and counterbalancing. *Psychol. Bull.*, **58**, 46–54.

Garside, R. F. 1956. The regression of gains upon initial scores. *Psychometrika*, **21**, 67–77.

Geisser, S., and Greenhouse, S. W. 1958. Extension of Box's results on the use of the *F* distribution in multivariate analysis. *Ann. math. Statist.*, **29**, 885–891.

Grant, D. A. 1956. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, **53**, 141–154.

———, 1962. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychol. Rev.*, **69**, 54–61.

Green, B. F., Jr., and Tukey, J. W. 1960. Complex analyses of variance: General problems. *Psychometrika*, **25**, 127–152.

Greenhouse, S. W., and Geisser, S. 1959. On methods in the analysis of profile data. *Psychometrika*, **24**, 95–112.

Grenander, U., and Rosenblatt, M. 1957. *Statistical analysis of stationary time-series.* New York: Wiley.

Guilford, J. P. 1954. *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. 1950. *Theory of mental tests.* New York: Wiley.

Guttman, L. 1940. Multiple rectilinear prediction and the resolution into components. *Psychometrika*, **5**, 75–99.

———, 1953. Image theory for the structure of quantitative variates. *Psychometrika*, **18**, 277–296.

———, 1954a. A new approach to factor analysis: The radex. In Lazarsfeld, P. F. (Ed.), *Mathematical thinking in the social sciences.* Glencoe, Illinois: The Free Press.

———, 1954b. Some necessary conditions for common-factor analysis. *Psychometrika*, **19**, 149–162.

———, 1955. The existence of total least-squares images. Unpublished manuscript.

———, 1956. "Best possible" systematic estimates of communalities. *Psychometrika*, **21**, 273–286.

———, 1957. *Communalities that maximize determinacy.* Berkeley: University of California Res. Rep. 16, Contr. No. AF 41(657)-76.

Haggard, E. A. 1957. Socialization, personality, and academic achievement in gifted children. *Sch. Rev.*, **65**, 388–414.

———, 1958. *Intraclass correlation and the analysis of variance.* New York: Dryden.

Harman, H. 1960. *Modern factor analysis.* Chicago: University of Chicago Press.

Harris, C. W. 1955. Separation of data as a principle in factor analysis. *Psychometrika*, **20**, 23–28.

———, 1956. Relationships between two systems of factor analysis. *Psychometrika*, **21**, 185–190.

———, 1962. Some Rao-Guttman relationships. *Psychometrika*, **27**, 247–263.

Harter, H. L., and Lum, M. D. 1955. Partially hierarchical models in the analysis of variance. *WADC Rep.* No. 55-33, March.

Haverland, E. 1954. Applications of an analytical solution for proportional profiles rotation. Doctor's Thesis, University of Illinois, October.

Hempel, C. G. 1952. *Foundations of concept formation in empirical science.* Chicago: University of Chicago Press.

Holmberg, A. 1954. Adventures in culture change. In Spencer, R. F. (Ed.), *Method and perspective in anthropology.* Minneapolis: University of Minnesota Press.

Holtzman, W. H. 1962. Methodological issues in *P*-technique. *Psychol. Bull.*, **59**, 248–256.

Hood, W. C., and Koopmans, T. C. 1953. *Studies in econometric method.* Cowles Com. Monogr. No. 14. New York: Wiley.

Horst, P. 1961*a*. Generalized canonical correlations and their applications to experimental data. *J. clin. Psychol.*, Monogr. Supp. No. 14, 331–347.

———, 1961*b*. Relations among *m* sets of variables. *Psychometrika*, **26**, 129–149.

Hughes, H. M., and Danford, M. B. 1958. Repeated measurements designs assuming equal variances and covariances. *USAF Sch. Aviat. Med. Rep.*, No. 59-40.

Jordan, N. 1962. Letter. *Amer. Psychologist*, **17**, 205–206.

Kaiser, H. 1958*a*. *The best approximation of a common-factor space.* Berkeley: University of California Res. Rep. 25, Contr. No. AF 41(657)-76.

———, 1958*b*. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

———, 1961. A note on Guttman's lower bound for the number of common factors. *Brit. J. statist. Psychol.*, **14**, 1–2.

———, 1962. Scaling a simplex. *Psychometrika*, **27**, 155–162.

Kelley, T. L. 1948. *Fundamentals of statistics.* Cambridge: Harvard University Press.

Kempthorne, O. 1952. *The design and analysis of experiments.* New York: Wiley.

Kendall, M. G. 1948. *The advanced theory of statistics.* London: C. Griffin and Co.

———, 1961. Natural Law in the social sciences. *J. Roy. Statist. Soc.*, Ser. A, **124**, 1–16.

Kish, L. 1953. Selection of the sample. In Festinger, L., and Katz, D. (Eds.), *Research methods in the behavioral sciences.* New York: Dryden.

Lawley, D. N. 1940. The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinburgh*, **60**, 64–82.

Lazarsfeld, P. F. 1947. *The mutual effects of statistical variables.* New York: Bureau of Applied Social Research, Columbia University.

Lazarsfeld, P. F., Berelson, B. R., and Gaudet, H. 1944. *The people's choice.* New York: Duell, Sloan, and Pearce.

Levenson, B. 1955. *Panel analysis workbook.* New York: Planning Project for Advanced Training in Social Research, Columbia University.

Lewis, D. 1960. *Quantitative methods in psychology.* New York: McGraw-Hill.

Lindquist, E. F. 1947. Goodness of fit of trend curves and significance of trend differences. *Psychometrika*, **12**, 65–78.

———, 1953. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

Lipset, S. M., Lazarsfeld, P. F., Barton, A. H., and Linz, J. 1954. The psychology of voting: An analysis of political behavior. In Lindzey, G. (Ed.), *Handbook of social psychology.* Cambridge, Addison-Wesley.

Lord, F. M. 1955. Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, **20**, 1–22.

———, 1956. The measurement of growth. *Educ. psychol. Measmt.*, **16**, 421–437. See also Errata, *ibid.*, 1957, **17**, 452.

———, 1958. Further problems in the measurement of growth. *Educ. psychol. Measmt.*, **18**, 437–454.

———, 1959. Statistical inferences about true scores. *Psychometrika*, **24**, 1–17.

———, 1960. Large-sample covariance analysis when the control variable is fallible. *J. Amer. Statist. Ass.*, **55**, 309–321.

Lubin, A. 1957. *Some rank-order tests for trend in a set of correlated means.* Washington, D.C.: Walter Reed Army Institute of Research.

———, 1959. *On the repeated measurements design.* Washington, D.C.: Walter Reed Army Institute of Research.

———, 1961. The interpretation of significant interaction. *Educ. psychol. Measmt.*, **21**, 807–817.

———, 1962. Statistics. In Farnsworth, P. R., and McNemar, Q. (Eds.), *Annual review of psychology.* Palo Alto, California: Annual Reviews, Inc.

McGregor, J. R. 1960. An approximate test for serial correlation in polynomial regression. *Biometrika*, **47**, 111–119.

McNemar, Q. 1955. *Psychological statistics* (2nd ed.). New York: Wiley.

———, 1958. On growth measurement. *Educ. psychol. Measmt.*, **18**, 47–55.

Mefferd, R. B., Jr., Moran, L. J., and Kimble, J. P., Jr. 1958. *Use of a factor analytic technique in the analysis of long term repetitive measurements made upon a single schizophrenic patient.* Paper presented at a symposium on multivariate analysis of repeated measurements on the same individual, American Psychological Association, September 3, 1958.

Mefferd, R. B., Jr., Moran, L. J., and Kimble, J. P., Jr., 1960. Chlorpromazine-induced changes in blood constituents in schizophrenia. *Transactions of the fourth research conference on chemotherapy and psychiatry.* Washington, D.C.: Veteran's Administration.

Meyer, W. J., and Bendig, A. W. 1961. A longitudinal study of the Primary Mental Abilities Test. *J. educ. Psychol.*, **52**, 50–60.

Mintz, A. 1946. A re-examination of correlations between lynchings and economic indices. *J. abnorm. soc. Psychol.*, **41**, 154–160.

Mood, A. McK. 1950. *Introduction to the theory of statistics.* New York: McGraw-Hill.

Moran, L. J. 1959. *Repetitive psychological measures.* Austin: University of Texas, Hogg Foundation.

Moran, L. J., Mefferd, R. B., Jr., and Kimble, J. P., Jr. 1960. Standardization of psychometric and psychodiagnostic tests for daily measurements in psychopharmacological research. *Transactions of the fourth research conference on chemotherapy and psychiatry.* Washington, D.C.: Veteran's Administration.

Mosier, C. I. 1943. On the reliability of a weighted composite. *Psychometrika,* **8,** 161–168.

Myers, J. L. 1959. On the interaction of two scaled variables. *Psychol. Bull.,* **56,** 384–391.

Neuhaus, J. O., and Wrigley, C. 1954. The quartimax method: An analytical approach to orthogonal simple structure. *Brit. J. statist. Psychol.,* **7,** 81–91.

Noble, M., Gruender, A., and Meyer, D. R. 1959. Conditioning in fish (*Mollienesia Sp.*) as a function of the interval between CS and US. *J. comp. physiol. Psychol.,* **52,** 236–239.

Orcutt, G. H., and James, S. F. 1948. Testing the significance of correlations between time series. *Biometrika,* **35,** 397–413.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. 1957. *The measurement of meaning.* Urbana: University of Illinois Press.

Plackett, R. L. 1960. Models in the analysis of variance. *J. Roy. Statist. Soc., Ser. B,* **22,** 195–217.

Potthoff, R. F., and Roy, S. N. 1962. *A generalized MANOVA model useful especially for growth curve problems.* Chapel Hill: University of North Carolina, Inst. Statist., Mimeo. Ser.

Quenouille, M. H. 1952. *Associated measurements.* New York: Academic Press.

———. 1957. *The analysis of multiple time series.* New York: Hafner.

Rankin, R. E., and Campbell, D. T. 1955. Galvanic skin response to negro and white experimenters. *J. abnorm. soc. Psychol.,* **51,** 30–33.

Rao, C. R. 1955. Estimation and tests of significance in factor analysis. *Psychometrika,* **20,** 93–111.

Rice, S. A. 1931. Experimental determinations by S. Wyatt and J. H. Fraser of the effects of rest pauses upon repetitive work. In Rice, S. A. (Ed.), *Methods in social science.* Chicago: University of Chicago Press.

Robson, D. S. 1959. A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced. *Biometrics,* **15,** 187–191.

Roy, J. 1958. Step-down procedure in multivariate analysis. *Ann. math. Statist.,* **29,** 1177–1187.

Roy, S. N., and Bargmann, R. E. 1958. Tests of multiple independence and the associated confidence bounds. *Ann. math. Statist.,* **29,** 491–503.

Roy, S. N., and Gnanadesikan, R. 1959. Some contributions to anova in one or more dimensions: I and II. *Ann. math. Statist.,* **30,** 304–317, 318–340.

Scheffé, H. 1959. *The analysis of variance.* New York: Wiley.

Shiskin, J., and Eisenpress, H. 1957. Seasonal adjustments by electronic computer methods. *J. Amer. Statist. Ass.,* **52,** 415–449.

Snedecor, G. W. 1946. *Statistical methods.* Ames: Iowa State College Press.

Stanley, J. C. 1956a. Fixed, random, and mixed models in the analysis of variance as special cases of finite model III. *Psychol. Rep.,* **2,** 369.

———. 1956b. Review of F. G. Cornell's "The essentials of educational statistics." *Educ. psychol. Measmt.,* **16,** 549–554.

Steel, R. G. D., and Torrie, J. H. 1960. *Principles and procedures of statistics.* New York: McGraw-Hill.

Thomson, G. H. 1924. A formula to correct for the effect of errors of measurement on the correlation of initial values with gains. *J. exp. Psychol.,* **7,** 321–324.

———. 1939. *The factorial analysis of human ability.* Boston: Houghton Mifflin.

———. 1951. *The factorial analysis of human ability* (5th ed.). Boston: Houghton Mifflin.

Thomson, G. H., and Ledermann, W. 1939. The influence of multivariate selection on the factorial analysis of ability. *Brit. J. Psychol.,* **29,** 288–306.

Thorndike, E. L. 1924. The influence of chance imperfections of measures upon the relation of initial score to gain or loss. *J. exp. Psychol.,* **7,** 225–232.

Thorndike, E. L., and Ruger, G. J. 1923. The effect of first year Latin upon knowledge of English words of Latin derivation. *School and Society,* **17,** 260–270, 417–418.

Thorndike, R. L. 1942. Regression fallacies in the matched groups experiment. *Psychometrika,* **7,** 85–102.

Tucker, L. R. 1959. Determination of generalized learning curves by factor analysis. In DuBois, P. H., *et al., Factor analysis and related techniques in the study of learning.* St. Louis: Washington University, Tech. Rep. No. 7, O.N.R. Contract No. Nonr-816(02).

Tukey, J. W. 1949. *Interaction in a row by column design.* Princeton: Princeton University, Memo Rep. 18.

Vernon, H. M., Bedford, T., and Wyatt, S. 1924. *Two studies of rest pauses in industry.* Med. Res. Council, Indus. Fatigue Res. Bd., Rep. No. 25. London: His Majesty's Stationery Office.

Walker, H. M., and Lev, J. 1953. *Statistical inference.* New York: Holt.

Webster, H. 1957. *Consistency statistics from sampling variances of test scores.* Poughkeepsie, New York: Vassar College, Mellon Foundation Research Report.

———. 1960a. A generalization of Kuder-Richardson reliability formula 21. *Educ. psychol. Measmt.,* **20,** 131–138.

———. 1960b. *Use of a general variance model for estimating test reliability.* Berkeley: University of California, Center for the Study of Higher Education.

———. 1962a. Item sampling as a sufficient but unnecessary requirement for precise mental testing. *Educ. psychol. Measmt.,* **22,** 321–324.

——, 1962b. Reliability of the ordering achieved by tests: A lower bound for generalized KR 21. *Psychol. Rep.*, **10**, 59-62.

Webster, H., and Abbott, W. 1956. *Forming groups that are homogeneous with respect to a large number of measures.* Poughkeepsie, New York: Vassar College, Mellon Foundation Research Report.

Whittle, P. 1951. *Hypothesis testing in time series analysis.* Uppsala: Almqvist.

Wilk, M. B., and Kempthorne, O. 1955. Fixed, mixed, and random models in the analysis of variance. *J. Amer. Statist. Ass.*, **50**, 1144-1167.

Wilk, M. B., and Kempthorne, O. 1957. Non-additivities in a Latin square design. *J. Amer. Statist. Ass.*, **52**, 218-236.

Williams, J. R. 1959. A test of the validity of the *P*-technique in the measurement of internal conflict. *J. Pers.*, **27**, 418-437.

Windle, C. 1954. Test-retest effect on personality questionnaires. *Educ. psychol. Measmt.*, **14**, 617-633.

Wiseman, S., and Wrigley, J. 1953. The comparative effects of coaching and practice on the results of verbal intelligence tests. *Brit. J. Psychol.*, **44**, 83-94.

Wishart, J., and Metakides, T. 1953. Orthogonal polynomial fitting. *Biometrika*, **40**, 361-369.

Wold, H. 1953. *Demand analysis.* New York: Wiley.

——, 1954. Causality and econometrics. *Econometrica*, **22**, 162.

——, 1956. Causal inference from observational data. A review of ends and means. *J. Roy. Statist. Soc., Ser. A*, **119**, 28-61.

Woodrow, H. 1932. Quotidian variability. *Psychol. Rev.*, **39**, 245-256.

——, 1938a. Factors in improvement with practice. *J. Psychol.*, **7**, 55-70.

——, 1938b. The relation of verbal ability to improvement with practice in verbal tests. *J. educ. Psychol.*, **29**, 179-186.

Wyatt, S., and Fraser, J. A. 1925. *Studies in repetitive work, with special reference to rest pauses.* Med. Res. Council, Indus. Fatigue Res. Bd., Rep. No. 32. London: His Majesty's Stationery Office.

Yule, G. U. 1926. Why do we sometimes get nonsense correlations between time series. *J. Roy. Statist. Soc.*, **89**, 1-64.

Yule, G. U., and Kendall, M. G. 1950. *An introduction to the theory of statistics* (14th ed.). New York: Hafner.

Zieve, L. 1940. Note on the correlation of initial scores with gain. *J. educ. Psychol.*, **31**, 391-394.

NORTHWESTERN UNIVERSITY

AN EXPLORATION OF TWO QUASI-EXPERIMENTAL DESIGNS: THE CROSS-LAGGED PANEL CORRELATION AND THE MULTIPLE TIME SERIES

A THESIS
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
for the degree
MASTER OF SCIENCE

Field of Psychology

By

RICHARD MARTIN ROZELLE

Evanston, Illinois

June 1965

Acknowledgments

The preparation of this thesis was supported by funds from the United States Office of Education, Education Media Branch, Project C-998, Contract 3-20-001, under the directorship of Professor Donald T. Campbell.

The author wishes to express gratitude in particular to Professor Donald T. Campbell, whose guidance and support made this study possible. The far-reaching intellectual influence of Professor Campbell is evidenced throughout the text.

Sincere appreciation is extended to Dr. Albert Erlebacher and Dr. Claude Mathis for their valuable criticisms and suggestions.

Grateful appreciation is extended to Mr. Woodrow Zoros and the Evanston Township High School, Evanston, Illinois, for providing the data used for analysis in this study.

The assistance given by the Northwestern Computing Center was invaluable in carrying out the computations necessary for the completion of this thesis.

Thanks are extended to Mrs. Irene Nolte for her helpful suggestions and conscientious proofreading of the manuscript.

Table of Contents

Page

List of Tables iv

List of Figures v

Introduction 1

Method 5

  Cross-Lagged Panel Correlation 5

  Multiple Time-Series Design. 6

  Data 9

Results. 16

  Analyses and Results for the Cross-Lagged Panel Correlations . . 16

  Analyses and Results for the Multiple Time-Series Design 30

Discussion 48

Summary. 52

References 54

Tables

Table Page

1. Experimental Design for the Time-Series Experiment Investigating the Effect of Absenteeism on Grades. 10
2. Means and Standard Deviations of Grades for Each Course Group and All-Course Data 14
3. Means and Standard Deviations of Absences for Each Course Group and All-Course Data 15
4. Cross-Lagged Panel Correlations for First Year Foreign Language (N = 1087). 17
5. Cross-Lagged Panel Correlations for Second and Third Year Foreign Languages (N = 911) 22
6. Cross-Lagged Panel Correlations for Combined Studies, English, and Social Studies (N = 5972) 24
7. Cross-Lagged Panel Correlations for Science Courses (N = 1612) 26
8. Cross-Lagged Panel Correlations for Mathematics (N = 2748) 27
9. Cross-Lagged Panel Correlations for the All-Course Data (N = 12,330) 29
10. The Significance of the Effect of Absence on Grades as Determined by the Mood Test Applied to the Multiple Time-Series Data 39
11. The t Test Between Means for the All-Course Data in the Multiple Time-Series Experiment. 42
12. The Chi-Square Test for the Composite p Value Obtained from the All-Course Data 45

Figures

Figure Page

1. The effect of absences on grades 8

2. Predicted effects of high absenteeism in a single marking period on grades 11

3. The effect of absences on grades for first year foreign language courses 31

4. The effect of absences on grades for second and third year foreign language courses. 32

5. The effect of absences on grades for combined studies, English, and social studies courses. 33

6. The effect of absences on grades for science courses 34

7. The effect of absences on grades for mathematics courses . . . 35

8. The effect of absences on grades for all courses combined . . 36

An Exploration of Two Quasi-Experimental Designs: The Cross-Lagged Panel Correlation and The Multiple Time Series

Richard M. Roselle

Northwestern University

The purpose of this study is the exploration of two quasi-experimental designs as applied to the retrospective analysis of archival data collected for administrative purposes. The program of quasi-experimental analysis seeks to extend into natural field situations the orientation of experimental analysis by the delineation of cause-and-effect relationships. Studies are classified as quasi-experimental where random assignment to treatments is not possible. Campbell and Stanley (1963) have assembled a considerable number of such designs, some of which have never been systematically used. Such unused designs need to be tested in application, for as any laboratory procedure, in the end we select those that work, those that lead to interpretable results.

The two designs chosen for exploration in this study are the Cross-Lagged Panel Correlation and The Multiple Time Series. These designs have been applied to the extensive records of a suburban high school in an effort to elucidate any causal relationship between absenteeism and grades.

Previous studies investigating the relationship between grades and absenteeism have produced conflicting results. A number of researchers have found a definite relationship between absenteeism and grades. Data collected by Butler (1925) indicated that secondary school students who received high marks were absent less frequently than students who received low marks. Although Wetzel (1928) did not describe his measure of ability nor give the size of his sample, he reported that senior high school students who were never absent or tardy surpassed the average of the total student body in achievement. Those students who were absent or tardy ten or more times during the school year were, at most ability levels, shown to be below the general average in achievement. Crider (1929) found that as the number of absences increased, the average number of grade points decreased. Studies carried out by Feingold (1928) and Jones (1931) indicate that the superior student had fewer absences than the poor student. Sandon (1938) examined academic marks of 406 male secondary students in the term following recorded absences. He found that the quality of work was definitely affected by the number of absences during the previous term. Young (1947) compared truants and nontruants with respect to achievement in arithmetic and English. There was no statistical evidence that truants were poorer in arithmetic than nontruants. However, there was a marked truant inferiority in English (reading and spelling). Tyerman (1958) discovered that truants were significantly inferior to their controls in both reading and arithmetic. Hersov (1960) discovered that the general standard of school work of truants was relatively poor. In a study carried out by Edmiston, Hinton, and Rasor (1947), a control group of pupils who had no absences during a semester was compared with a group who had ten or more unexcused absences. The average grades of the two groups were found to be significantly different. Anikeef (1954), using 2654 college students found a statistically significant inverse relationship between the number of class absences and grade quality points. This relationship was significant for all class levels (freshmen through senior). Amthauer (1963) compared intelligence, performance, and teacher ratings for high absentee and low absentee groups of apprentices. He found that the low absentee group scored better in all categories than the high absentee group.

On the other hand, a number of studies have found no relationship. Heilman (1928) using 828 eleven-year-old school children found no significant relationship between grades and attendance. Similar results were obtained from 157 elementary school children in a study conducted by Woolf (1926). Odell (1923), used correlations to study the relationship between absenteeism in one semester and grades in the following semester for students in the first through eighth grades. He also correlated grades and attendance records within the same semester. His results indicated very little relation between absenteeism and grades within the same semester and even less relationship between prior absenteeism and subsequent (in the following semester) grades. In the latter case most of the correlations were below .100 with an over-all correlation of .08. Jensen (1930) found a negligible correlation between attendance and achievement in high school algebra for the 80 cases on which he obtained data. Other investigations summarized by Stephens (1933) lead to the conclusion: "The data uniformly indicate that any variations in the amount of attendance have had surprisingly small effects" on achievement. Nemzek (1938) criticized the studies summarized by Stephens, indicating that they could not be considered conclusive. The reasons given were that there was only slight variability in attendance in the groups studied, the amount of time was limited to only one or two semesters, and age, sex, race, and other factors had not been singled out. A study which attempted to control for these factors was carried out by Nemzek (1938). The results of this study indicated that an increase in the amount of absence was generally accompanied by a slight decrease in school marks. However, the amount of absence had a negligible effect for purposes of differential prediction. Almost identical results were obtained in another study by Finch and Nemzek (1935). Cutler (1936) working with third grade pupils found that when IQ was held constant the relation between attendance and achievement while positive, was too low to be significant. Similar results were obtained with college in a study conducted by Nelson (1937). Using 300 undergraduates of the University of Minnesota High School, Finch and Nemzek (1940) discovered no relationship between absenteeism and grades for boys, and only a correlation of +.3 for girls. Platt (1943), pooling boys and girls, found a correlation of +.3 between school attendance and scholastic points. In an experiment comparing grades of college students who attended a psychology class with compulsory attendance and those who registered for a similar class with no attendance required, which resulted in many more absences, Caro (1962) found no significant difference.

Considering the variety of results displayed by the above review, the relationship between absence and grades remains unclear. While many studies have found the expected relationship, many have not. It is perhaps comforting that none have found more absences to go with better grades. Probably a low but significant relationship between absences and grades would be found if all studies could be averaged. But even here, the direction of effect would be only presumed, not demonstrated. The correlation could be explained as well by assuming that low grades caused greater absenteeism as vice versa.

The present study brings previously unused methods to this setting in an attempt to explicate the causal relationship. The major focus is on the exploration of the methods.

Method

Cross-Lagged Panel Correlation

Cross-Lagged Panel Correlations are correlations between measures of variables taken at different times for the purpose of comparing the relative magnitude of the antecedent-subsequent relationships that exist. The principal reason for using Cross-Lagged Panel Correlations is to aid in deciding upon the direction of causation in the situation where two series of data correlate. If C is cause and E is effect, then $r_{C_n E_{n+1}} > r_{C_{n+1} E_n}$ with *n* designating the relative time the measurement is taken (Campbell, 1963, p. 236). In the present case the question is asked, "In secondary school, does poor attendance cause subsequent lower grades or do low grades cause subsequent poor attendance?" It is felt that previous studies have, in general, shown that some relationship does exist between grades and attendance. Although the causal relations are probably in both directions, it would seem that, intuitively, the greater relative cause would be in the direction from poor attendance to low grades, particularly in college preparatory subjects. In Cross-Lagged Panel Correlation notation, the general prediction is that $r_{A_1 G_2} > r_{A_2 G_1}$ where A represents attendance and G represents grade and the numbered subscripts represent the relative time the measurement was taken.

In the present case, the Cross-Lagged Panel Correlation matrix contains 18 variables. The important distinction between these variables is that they represent measurements of grades and attendance at different times throughout the nine-month period. There are eight attendance records and ten grade records for each student. The time-lagged Correlations between these variables should reveal an index of relative preponderance and therefore the direction in which a causal inference may be made.

It is expected that the effect of a high absence in one marking period will be the greatest on the grade of the marking period immediately following. This effect should diminish the further the marking periods are apart from each other in time. Thus, $r_{A_1O_2} > r_{A_1O_3} > r_{A_1O_4} > r_{A_1O_5} > r_{A_1O_6}$.

These effects should be maximized in those courses making least use of pre-course learnings, and where within the course there was most dependence upon cumulative knowledge so that the comprehension of what is taught first is needed to understand what is taught later. It was felt that this day to day dependent process of accumulating knowledge was best represented in first year foreign language courses. Since all of the students are being introduced to a new language, it seems safe to assume that each of these students initially possesses the same relatively low level of knowledge about the particular language being introduced. For comparative and exploratory purposes, other courses were also included in the analyses.

A large sample N is needed to balance out fluctuations in population measures and because the analysis depends upon the significance of small differences in correlation.

Multiple Time-Series Design

A discussion of Time-Series Designs as a quasi-experimental method is presented in Campbell and Stanley (1963) and Campbell (1963). The Time-Series Experiment consists of periodic measurements of an individual or group of individuals. An experimental treatment is presented to the group at some time during the course of the periodic measurements. If the treatment has produced an effect, a discontinuity of measurements is recorded in the Time Series.

In the present study, "treatment" groups consist of members selected on the basis of having a minimal amount of absences in all but one of the marking periods and a maximum amount of absences in that one particular marking period. All students falling between the lower limit of the high absence criterion and the upper limit of the low absence criterion are not used for data analysis. Thus, the "treatment" is an unusually high amount of absences in one and only one marking period. A "control" group will consist of those students having none or a minimum amount of absences in all of the marking periods. The effect of high absence marking period upon the mean grade level of each treatment group is considered a measure of the treatment effect.

It is expected that the Time-Series Analysis will reveal a significant drop in the mean grade level for the marking period having the maximum amount of absences. The mean grade level of the marking periods following the "treatment marking period" should have a tendency to rise to the level maintained previous to the treatment. The "control" group should maintain about the same mean grade level throughout all of the marking periods with no significant changes in any one marking period. If high absenteeism does have a major effect on grades, it should produce a discontinuity of significant magnitude in the Time-Series measures beyond the random fluctuations of the low absence marking period measures. The results should be similar to the curves in Figure 1.

Since both pretest and posttest measurements are necessary for a Time-Series Analysis, four treatment groups and one control for each

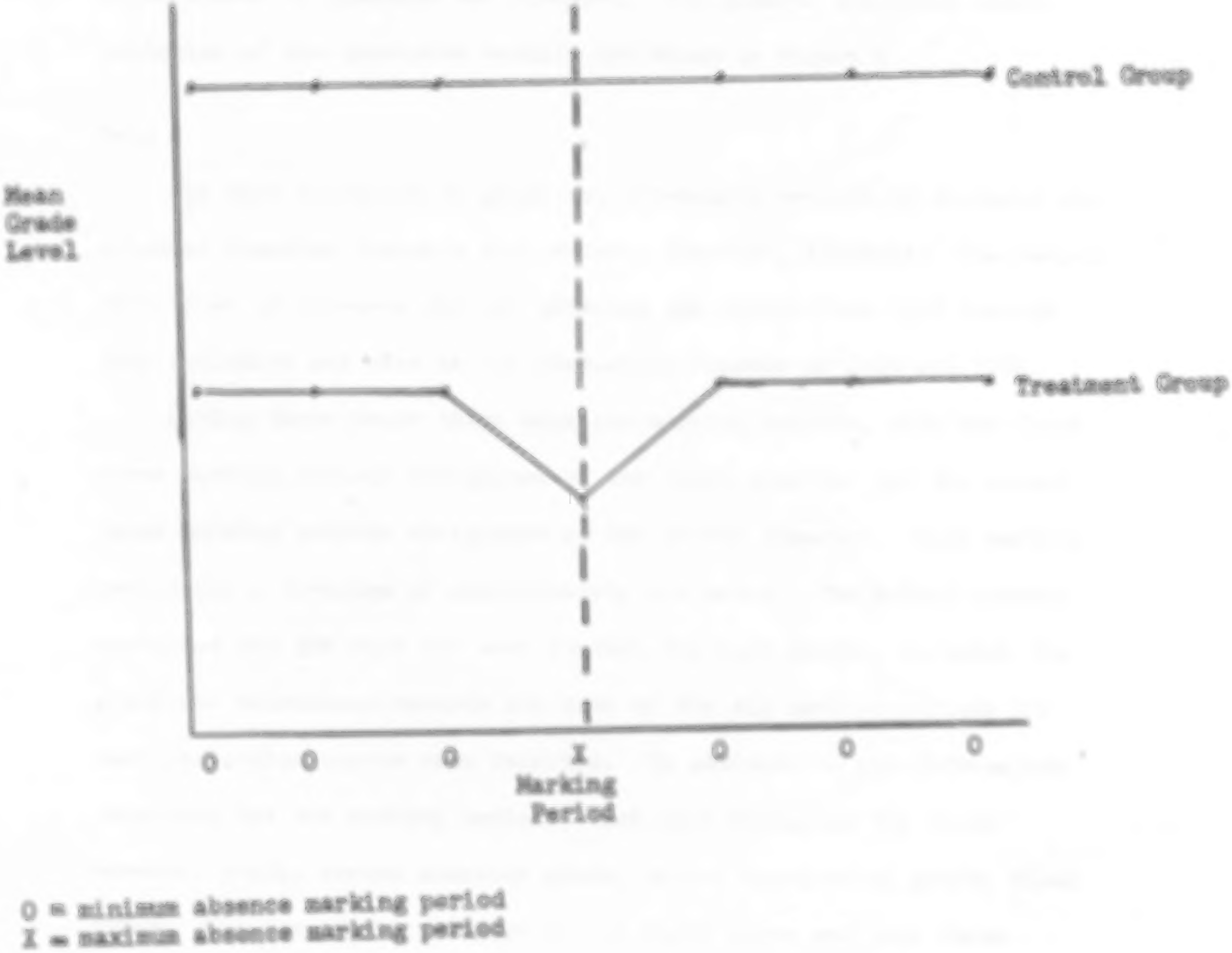


Fig. 1. The effect of absences on grades.

course categories were used. Table 1 shows the design used for this analysis.

The prediction was made that if absences are effective, the grades most affected would be those given for the marking period in which the major number of absences was recorded. The general graphical representation of the predicted results are shown in Figure 2.

Data

The data consisted of grade and attendance records of students who attended Evanston Township High School, Evanston, Illinois. The records were those of students who had attended the school from 1959 through 1962 inclusive and were in the graduating classes of 1963 and 1964.

During three years there were six marking periods, with the first three marking periods designated as the first semester and the second three marking periods designated as the second semester. Each marking period had a duration of approximately six weeks. The school records contained one IBM card for each student for each course, on which the grade and attendance records for each of the six marking periods for that particular course were recorded. In addition to the information regarding the six marking periods, each card contained the first semester grade, second semester grade, spring examination grade, final grade. The sum of the absences in the first three and last three marking periods were subsequently computed as a part of the present analysis. It is important to emphasize that the unit of analysis was the information given on each IBM card and not the individual student. The number of IBM cards was the N used in the statistical analyses, not the number of students. Because a number of different courses were used,

Table 1

Experimental Design for the Time-Series Experiment Investigating the Effect of Absenteeism on Grades

	Marking Period					
Treatment Group	1	2	3	4	5	6
2	0	X	0	0	0	0
3	0	0	X	0	0	0
4	0	0	0	X	0	0
5	0	0	0	0	X	0
Control	0	0	0	0	0	0

X = Major Absences (ten days or over)

0 = Minimal Absences (two days or under)

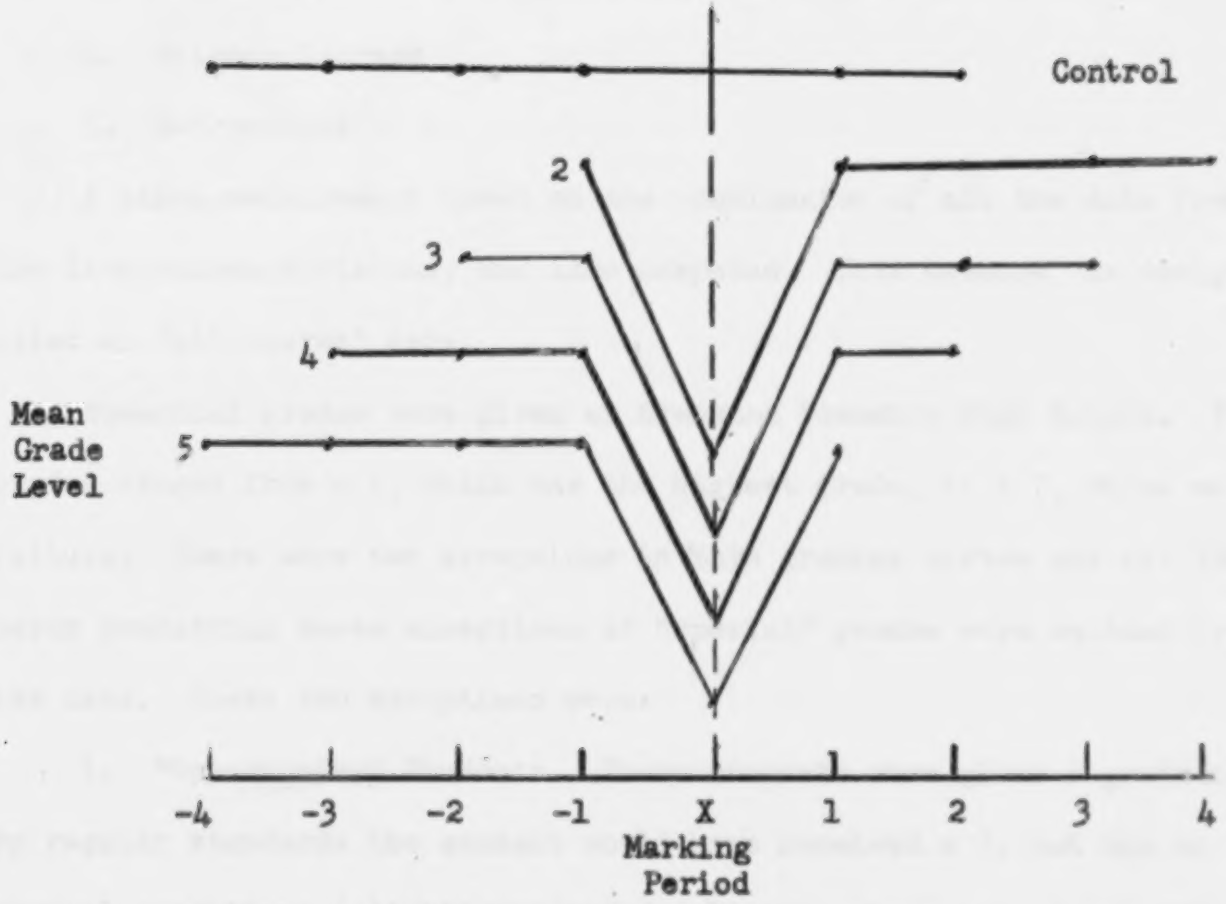


Fig. 2. Predicted effects of high absenteeism in a single marking period on grades.

most of the students were represented in more than one analysis, and represented more than once in the pooled analysis.

A total of five different course combinations offered by the Evanston Township High School were analyzed. These were:

1. First Year Foreign Languages
2. Second and Third Year Foreign Languages
3. Combined Studies, English, and Social Studies
4. Science Courses
5. Mathematics

A sixth measurement based on the combination of all the data from the five course divisions, was also computed. This measure is designated as "all-course" data.

Numerical grades were given at Evanston Township High School. These grades ranged from a 1, which was the highest grade, to a 7, which was a failure. There were two exceptions in this grading system and all IBM cards containing these exceptions of "special" grades were omitted from the data. These two exceptions were:

1. "Opportunity" Students. These students were given a grade of 6. By regular standards the student would have received a 7, but due to special conditions (slightly retarded, slower workers, etc.) was not failed.

2. "Basic" Students. These students received a B instead of a numerical grade. The B classification indicated that due to certain criterion factors established by the school, the student could not receive a grade higher than 4 (i.e., his grade range was from 4 to 7).

It should be noted that due to the particular grading system used, the predicted correlations among grades and absences will result in a positive relationship between the two variables. In other words, the "smaller" the grade, the fewer the number of absences that is expected. The "larger" grades, because they are lower, should have a correspondingly greater number of absences associated with them.

However, a problem exists concerning the normality of the sampling distribution of the correlation among grades and absences: the distribution of absences is markedly skewed. Because of the nature of the "absence" variable, its sampling distribution contains a mode of zero and neither a logarithmic nor any other type of transformation can remove this mode around zero. In a normal school setting such as that of the Evanston Township High School, it simply occurs that a great number of the students have no absences in any six week marking period. Thus, because of the impossibility of normalizing the sampling distribution of absences, no attempt was made to reduce its skewness.

The mean numerical grades and number of absences are given for each course division and the all-course data in Tables 2 and 3. Beneath each mean is listed the standard deviation for that variable. It is seen that a similar pattern of means and standard deviations exists for each course grouping. The two largest means and corresponding standard deviations appear where they would most likely be expected: the sum of the absences for the first three and last three marking periods.

Table 2

Means and Standard Deviations of Grades for Each Course Group and All-Course Data

		G_1	G_2	G_3	G_4	G_5	G_6	G_{123}	G_{456}
First Yr. Foreign Languages	Mean	2.5805	2.5649	2.6743	2.7654	2.6946	2.6670	2.6035	2.6624
	Standard Deviation	1.2284	1.1286	1.1574	1.1528	1.1477	1.1656	1.1075	1.1063
Second & Third Yr. Foreign Languages	Mean	2.7640	2.6487	2.6696	2.7881	2.6926	2.6531	2.6575	2.6608
	Standard Deviation	1.1214	1.1171	1.1072	1.1407	1.1578	1.2200	1.0562	1.1282
Combined Studies, English & Social Studies	Mean	3.1328	2.9899	2.9956	3.0630	2.9926	2.9494	3.0228	2.9642
	Standard Deviation	1.0618	1.1210	1.1529	1.1616	1.1608	1.1980	1.0600	1.1146
Science	Mean	2.8145	2.8251	2.7624	2.8232	2.8244	2.7049	2.7674	2.7215
	Standard Deviation	1.2371	1.2676	1.2850	1.2652	1.3092	1.2912	1.2077	1.2170
Mathematics	Mean	2.6881	2.5830	2.7300	2.7132	2.6488	2.6761	2.6619	2.6492
	Standard Deviation	1.1409	1.1774	1.2094	1.2142	1.2013	1.2940	1.1071	1.1903
All-Course	Mean	2.9161	2.8150	2.8535	2.9071	2.8456	2.8097	2.8450	2.8132
	Standard Deviation	1.1441	1.1688	1.1889	1.1950	1.1983	1.2398	1.1088	1.1552

Table 3

Means and Standard Deviations of Absences for Each Course Group and All-Course Data

		A_1	A_2	A_3	A_4	A_5	A_6	A_{123}	A_{456}
First Yr. Foreign Languages	**Mean**	1.4066	1.1389	1.7415	1.7479	1.3293	1.5179	4.2870	4.5942
	Standard Deviation	2.6241	2.0843	2.6343	2.5408	2.2332	2.4997	5.8060	5.7888
Second & Third Yr. Foreign Languages	**Mean**	1.5587	1.1109	1.9670	1.6948	1.3074	1.6081	4.6367	4.6103
	Standard Deviation	1.8435	1.8822	2.3662	2.3699	1.7384	2.6343	4.3640	4.9234
Combined Studies English & Social Studies	**Mean**	1.3026	1.0830	1.7200	1.5740	1.2096	1.4077	4.1057	4.1914
	Standard Deviation	2.0706	1.8794	2.5693	2.4180	1.9661	2.2613	4.9357	5.1145
Science	**Mean**	1.3685	1.1166	1.9181	1.6669	1.2630	1.4690	4.4032	4.3990
	Standard Deviation	2.1474	1.8039	2.7307	2.3895	1.9309	1.3003	4.6929	4.9341
Mathematics	**Mean**	1.4290	1.1226	1.8941	1.6073	1.3086	1.4610	4.4458	4.3770
	Standard Deviation	2.1514	1.9205	2.5999	2.3627	1.9579	2.2037	5.1749	4.8563
All-Course	**Mean**	1.3675	1.1032	1.8049	1.6178	1.2564	1.4521	4.2756	4.3264
	Standard Deviation	2.1530	1.8979	2.5906	2.4100	1.9693	2.3102	5.0050	5.0860

Results

Analyses and Results for the Cross-Lagged Panel Correlations

In terms of a priori anticipation, the clearest effect was expected from first year foreign languages. Table 4 presents Cross-Lagged Panel Correlations for absences and grades in the first year foreign language courses, plus other data providing interpretative background. These include the time 1-time 2 correlations for the same variables, test-retest reliabilities as they might be called, and the cross-variable correlations within the same time period, i.e., unlagged or syncronous correlations. Noteworthy are the very low values of the correlations between grades and absences, both lagged and unlagged. In spite of these very low values, in all cases the direction of lagging favors the absence-causes-poor grades interpretation. Are these slight differences statistically significant?

The problem is a complicated one on a number of scores. First of all, as discussed earlier, there is the skewness of the absences with the modal value at zero. This fact makes technically inapplicable the normal curve assumptions of the tests of significance. Lacking anything better, however, they have been used. The significance of difference between two correlation coefficients also presents unsolved problems. When the two coefficients are computed on the same population, and the arrays of the coefficients involved are correlated--even if none are repeated--then this correlation of arrays reduces the sampling variation and should be taken into account in the error term. Peters and Van Voorhis (1940, p. 185) provide a Person and Filon test for this situation, as well as the complex procedure for computing the correlation between correlation

Table 4

Cross-Lagged Panel Correlations for First Year Foreign Language (N = 1087)

	Test-Retest		Synchronous Cross-Correlations		Cross-Lagged Correlations		Sign	t_r	Single Test p	All Test p	t_z'	Single Test p	All Test p
	Grades	Absences	T_1	T_2	A_1G_2	A_2G_1	$A_1G_2 > A_2G_1$						
Adjacent Periods													
1-2	.780	.389	.003	.018	.008	-.028	+	1.00			.84		
2-3	.782	.523	.018	.106	.039	.001	+	1.16			.89		
3-4	.790	.402	.106	.035	.103	-.038	+	4.01	.001	.01	3.29	.01	
4-5	.789	.514	.035	.017	.025	-.006	+	.94			.72		
5-6	.797	.469	.017	.116	.041	.014	+	.79			.63		
Non Adjacent													
1-3	.678	.391	.003	.106	.015	-.029	+	1.19			1.02		
2-4	.724	.396	.018	.035	.042	-.052	+	2.60	.01		2.60	.05	
3-5	.733	.395	.106	.017	.074	-.011	+	2.36	.05		1.98		
4-6	.766	.374	.035	.116	.026	-.009	+	.47			.40		
1-4	.652	.257	.003	.035	.042	-.080	+	3.13	.01		2.84	.01	
2-5	.687	.382	.018	.017	.042	-.038	+	2.18	.05		1.86		
3-6	.701	.360	.106	.116	.112	-.019	+	3.56	.001	.05	3.06	.01	
1-5	.594	.278	.003	.017	.021	-.084	+	2.69	.01		2.45	.05	
2-6	.646	.323	.018	.116	.075	-.035	+	2.89	.01		2.56	.05	
1-6	.545	.244	.003	.116	.051	-.051	+	2.56	.01		2.38	.05	
Semester 1,2,3-4,5,6	.752	.532	.014	.035	.081	-.063	+	4.36	.001	.01	3.36	.01	

coefficients used to reduce the error term. This involves all of the other correlations among the four variables presented in the first four columns of Table 4 designated as t_r.

More orthodox today is the use of Fisher's z' transformation of correlation coefficients in factors and their significance. No z' formula reducing the error term for the correlation between correlation coefficients is available. These run consistently slightly smaller than do the other t's, due to the larger error term.

Were only one t value to be examined for significance, or were all t values uniformly high, these significance levels would seem appropriate. Where, however, many tests are computed and only a few reach such significance levels, even these few may be suspect of being chance occurrences made more probable by the many tests involved. In such an instance, a more conservative revision of the significance level can be computed by asking how often would a value of this magnitude occur by chance in 100 sets of N tests. the $p < .01$ level becomes the $\frac{1}{100 \times N}$ normal curve table. For the $p < .05$ level, the new value is $\frac{1}{20 \times N}$. There was a total of 15 Cross-Lagged-Panel comparisons representing single marking periods in each of the five course divisions. With the z values obtained from Fisher and Yates (1953, p. 59) Table VIII4, the adjusted significance levels are:

(a) $\frac{1}{100 \times 15 \times 5}$ or $p = \frac{.000133}{2} = .000066667$ (two-tailed test),

$$z = 3.82 \text{ for } p < .01$$

(b) $\frac{1}{20 \times 15 \times 5}$ or $p = \frac{.000667}{2} = .0003334$ (two-tailed test),

$$z = 3.40 \text{ for } p < .05$$

These adjusted or whole set significance levels are indicated in Table 4 as "all-test p" values. Ryan (1960) has distinguished between error rate per experiment and error rate experimentwise. Error rate per experiment is defined as the number of comparisons incorrectly called significant, divided by the total number of experiments. This error rate is tested by the "all-test p" values. Error rate experimentwise is defined as the number of experiments containing erroneous statements of significance, divided by the total number of experiments. Ryon states that the error rate experimentwise can never be greater than the error rate per experiment because there cannot be more experiments with erroneous conclusions than there are errors in individual comparisons. The error rate experimentwise was computed by the following formula:

$$1 - (1-x)^k = .05 \qquad (.05 \text{ level})$$

$$1 - (1-x)^k = .01 \qquad (.01 \text{ level})$$

For both levels, $k = 75$, which was the number of time-lagged comparisons made for the five course divisions. The z values obtained from this method were, $z = 3.82$ for $p < .01$ and $z = 3.40$ for $p < .05$. These values are the same as those obtained from the error rate per experiment method. Thus, with k equal to 75, the error rate experimentwise technique does not produce a significantly less stringent level for testing internal differences. For the purposes of this study, then, the distinction between error rate per experiment and error rate experimentwise is not important in that it has no differentiating effect upon the statistical interpretations of the comparisons made. The regular or unconservative significance levels ($z = 2.57$ for $p < .01$, $z = 1.96$ for $p < .05$) are indicated as "single-test p" values in Table 2. Only three of the t_r values are significant when the "all-test p" criterion is used whereas ten of these values are significant in reference to the "single-test p." The uniform direction of these values, and their occasionally high significance, especially for the semester values, seems to provide confirmation of the predicted relationship. However, two further considerations weaken this picture.

It is observed that the prediction of decreasing differences between the time-lagged correlations as the time lag increased is not confirmed by the results shown in Table 4. Of the nine "single-test p" values indicating the significance of t_r for individual marking periods, eight occur for nonadjacent marking periods. In fact, these results are almost opposite to the prediction; Table 4 shows that as the time lag between marking periods decreased, fewer and fewer of the t_r values were significant. Although the "all-test p" values that indicated t_r significant did not show a relationship opposite to the prediction, neither did they confirm the prediction. Of the two "all-test p" values that indicated significance, only one occurred for adjacent working periods, the other recurring for the nonadjacent third and sixth marking periods.

Perhaps even more important is the persistent trend towards negative correlations in the $r_{A_2G_1}$ column of Table 4. This sign to the values was unexpected in the causal hypotheses involved. However, it is entirely possible that prior good grades could either (1) cause the student to become more conscientious and subsequently attend school more regularly, indicating a positive relationship between G_1 and A_{n+1} or, (2) cause the student to become more confident and reduce to him the importance of class attendance, indicating a negative relationship between G_1 and A_{n+1}. Low or "bad" grades could either cause the student to (1) become frustrated and reduce his initiative to work harder thus producing more absences and resulting in a positive relationship, or (2) cause the student to attempt to improve his grades, thus improving his attendance record and producing a negative relationship between grades and absences. For either high grades or low grades or both, the second alternative effects indicating negative relationships seem to have occurred in the data. Such a process works to produce the same direction of differences as does the high absence low grade model. In the present case, the absolute sign of the correlation coefficients points to this possibility, but if the causal effect were superimposed upon a generally positive level of cross-correlation due to other factors, then the two codirectional hypotheses would be unextricable. This ambiguity is a major weakness of the Cross-Lagged Panel Correlation not noted in previous reports.

Table 5 presents Cross-Lagged Panel Correlations for the second and third year foreign language courses plus other information identical to that explained for Table 4. As was the case in Table 4, the correlations between grades and absences are extremely low. In all but four cases the direction of the lagging agrees with the absences-cause-poorer-grades prediction. None of the $\underline{t}_r$ values in Table 5 are significant for the "all-test $\underline{p}$" criterion. Only four $\underline{t}_r$ values are significant at the "single-test $\underline{p}$" level. In none of the four cases in which the direction of the lagging did not agree with the absences-cause-poorer-grades prediction (i.e., negative sign) were the $\underline{p}$ values significant. As in Table 4 the difference between the Cross-Lagged Panel Correlations for first and second semester grades and absences was significant and in the predicted direction. Due to the small magnitude of the Cross-Lagged-Panel Correlations and the lack of significant differences between them,

Table 5

Cross-Lagged Panel Correlations for Second and Third Year Foreign Languages (N = 911)

	Test-Retest		Synchronous Cross-Correlations		Cross-Lagged Correlations		Sign	$\underline{t}_r$	Single Test	All Test	$\underline{t}_z'$	Single Test	All Test
									$\underline{p}$	$\underline{p}$		$\underline{p}$	$\underline{p}$
	Grades	Absences	T_1	T_2	A_1G_2	A_2G_1	A_1G_2 A_2G_1						
Adjacent Periods													
1-2	.737	.283	.108	.072	.053	.024	+	.70			.62		
2-3	.758	.247	.072	.112	.062	.075	-	-.31			-.28		
3-4	.789	.259	.112	.035	.098	-.009	+	2.57	.01		2.29	.05	
4-5	.774	.309	.035	.103	.005	.045	-	-.98			-.85		
5-6	.781	.306	.103	.050	.094	.017	+	1.89	.05		1.65		
Non-Adjacent Periods													
1-3	.715	.268	.108	.112	.030	.049	-	-.45			-.41		
2-4	.744	.242	.072	.035	.034	-.006	+	.94			.85		
3-5	.732	.304	.112	.103	.087	.049	+	.92			.81		
4-6	.760	.265	.035	.050	.014	.006	+	.19			.17		
1-4	.693	.215	.108	.035	.044	-.017	+	1.41			1.30		
2-5	.715	.243	.072	.103	.032	.025	+	.16			.15		
3-6	.731	.239	.112	.050	.102	-.012	+	2.69	.01		2.44	.05	
1-5	.669	.251	.108	.103	.029	.030	-	-.02			-.02		
2-6	.719	.151	.072	.050	.021	-.017	+	.86			.81		
1-6	.661	.175	.108	.050	.047	-.003	+	1.14			1.07		
Semester 1,2,3-4,5,6	.816	.438	.109	.058	.093	.014	+	2.11	.05		1.69		

no conclusions can be drawn concerning the prediction of the decreasing effect of absences upon grades as the time lag is increased.

The Cross-Lagged Panel Correlations for absences and grades for combined studies, English, and social studies are presented in Table 6 It is noted that most of the values for the correlations between absences and grades, both lagged and synchronous, are higher than the corresponding value in Table 4 and Table 5. However, Table 6 also shows that in most of the cases, the direction of the lags is not in the predicted direction. In eleven of sixteen lags, $r_{A_2G_1}$ has a larger value than $r_{A_1G_2}$ which favors a grades-cause-absences interpretation and which is contrary to the causal interpretation made in the predictions. Three $\underline{t}_r$ values are significant in Table 6 all at the "all-test $\underline{p}$" or conservative level, and all in a negative direction. It seems, then, for the combined studies, English, and social studies courses shown in Table 6, that the preponderance of cause lies with previous grades, not with previous absences. If this result is applied to the prediction dealing with increasing time lags, then Table 6 should show a decreasing effect of prior grades on subsequent absences as the time lag is increased. This is not the case, however. None of the significant $\underline{t}_r$ values occur for adjacent marking periods. In fact, two of the significant $\underline{t}_r$ values occur in the largest possible time lags: the first and fifth marking periods, and the first and sixth marking periods.

The results of Table 6 are indeed puzzling. Perhaps some characteristics specific to combined studies, English, and social studies courses at Evanston Township High School could account for the reversal in the prediction made concerning absences and grades. Whether this was the

Table 6

Cross-Lagged Panel Correlations for Combined Studies, English, and Social Studies (N = 5972)

	Test-Retest		Synchronous Cross-Variables		Cross-Lagged Correlations		Sign	t_r	Single Test	All Test	t_z'	Single Test	All Test
	Grades	Absences	T_1	T_2	A_1G_2	A_2G_1	$A_1G_2 > A_2G_1$		p	p		p	p
Adjacent Periods													
1-2	.766	.385	.088	.125	.060	.099	-	-2.55			-2.14	.05	
2-3	.755	.382	.125	.173	.130	.131	-	- .07			- .06		
3-4	.775	.377	.173	.127	.136	.105	+	2.04			1.72		
4-5	.778	.391	.127	.160	.106	.137	-	-2.05			-1.72		
5-6	.783	.428	.160	.162	.144	.129	+	1.02			.84		
Non-Adjacent Periods													
1-3	.734	.305	.088	.173	.069	.135	-	-4.13	.001	.01	-3.64	.001	.05
2-4	.744	.306	.125	.127	.107	.087	+	1.25			1.10		
3-5	.754	.344	.173	.160	.129	.123	+	.39			.33		
4-6	.764	.348	.127	.162	.119	.127	-	- .52			- .44		
1-4	.722	.238	.088	.127	.068	.087	-	-1.15			-1.04		
2-5	.738	.328	.125	.160	.107	.124	-	-1.08			- .94		
3-6	.760	.334	.173	.162	.125	.113	+	.77			.66		
1-5	.700	.310	.088	.160	.052	.116	-	-3.98	.001	.01	-3.52	.001	.05
2-6	.744	.325	.125	.162	.101	.102	-	- .06			- .06		
1-6	.713	.283	.088	.162	.047	.114	-	4.13	.001	.01	-3.69	.001	.05
Semester 1,2,3-4,5,6	.837	.546	.162	.188	.141	.154	-	- .98			- .73		

case or whether chance fluctuation in the data was the predominant factor, the absolute magnitude of correlation coefficients is so small that meaningful and conclusive interpretation is not possible at this time.

Table 7 presents the Cross-Lagged Panel Correlation information for science courses. The values of the lagged and synchronous correlations are larger than those in Tables 4 and 5 and approach more closely the size of the value found in Table 6. However, most of the lags in Table 7 are in the positive or predicted absences-cause-poorer-grades direction. In eleven out of sixteen cases, the lags are in a positive direction. Unfortunately, none of the differences between the Cross-Lagged Panel Correlations are significant at the "all-test p" level and only four t_r values are significant at the "single-test p" level. Because two of the significant t_r values occur in adjacent marking periods and two in nonadjacent marking periods, and because the values of the correlation coefficients are so small, no conclusions can be stated for the increasing lag-decreasing t prediction.

In Table 8 are the Cross-Lagged Panel Correlations for mathematics courses. None of the differences between the lagged correlations are significant at the "all-test p" level and only three are significant in reference to the "single-test p" level. In spite of the small differences, all but five of the lags favor the absences-cause-poorer-grades interpretation. Again, because of the lack of significant differences and the small size of the correlation coefficients, the hypothesis concerning the increasing lag and resultant decreasing effect of absences upon grades cannot be effectively evaluated.

Table 7

Cross-Lagged Panel Correlations for Science Courses (N = 1612)

	Test-Retest		Synchronous Cross-Variables		Cross-Lagged Correlations		Sign	t_r	Single Test	All Test	t_z'	Single Test	All Test
	Grades	Absences	T_1	T_2	A_1G_2	A_2G_1	A_1G_2 A_2G_1		p	p		p	p
Adjacent Periods													
1-2	.769	.258	.127	.170	.086	1.22	-	-1.15			-1.03		
2-3	.796	.228	.170	.147	.128	.093	+	1.11			1.01		
3-4	.789	.251	.147	.157	.135	.102	+	1.06			.95		
4-5	.796	.366	.157	.154	.096	.161	-	-2.22	.05		-1.88		
5-6	.806	.343	.154	.098	.132	.058	+	2.49	.05		2.12	.05	
Non-Adjacent Periods													
1-3	.769	.210	.127	.147	.104	.095	+	.28			.26		
2-4	.757	.321	.170	.157	.111	.096	+	.49			.43		
3-5	.767	.276	.147	.154	.107	.100	+	.23			.20		
4-6	.785	.267	.157	.098	.090	.067	+	.74			.66		
1-4	.702	.212	.127	.157	.089	.099	-	- .31			- .29		
2-5	.752	.283	.170	.154	.114	.126	-	- .39			- .34		
3-6	.776	.227	.147	.098	.109	.029	+	2.52	.05		2.28	.05	
1-5	.702	.231	.127	.154	.103	.105	-	- .06			- .06		
2-6	.762	.288	.170	.098	.122	.058	+	2.07	.05		1.83		
1-6	.698	.235	.127	.098	.110	.056	+	1.69			1.54		
Semester 1,2,3-4,5,6	.847	.490	.180	.157	.172	.125	+	1.77			1.36		

Table 8

Cross-Lagged Panel Correlations for Mathematics (N = 2748)

	Test-Retest		Synchronous Cross-Variables		Cross-Lagged Correlations		Sign	$\underline{t}_r$	Single Test	All Test	$\underline{t}_r'$	Single Test	All Test
									$\underline{p}$	$\underline{p}$		$\underline{p}$	$\underline{p}$
	Grades	Absences	T_1	T_2	A_1G_2	A_2G_1	A_1G_2 A_2G_1						
Adjacent Periods													
1-2	.718	.412	.070	.085	.073	.017	+	2.48	.05		2.08	.05	
2-3	.737	.376	.085	.142	.065	.092	-	-1.18			-1.01		
3-4	.750	.314	.142	.131	.136	.070	+	2.83	.01		2.47	.05	
4-5	.744	.332	.131	.180	.094	.145	-	-2.21	.05		-1.92		
5-6	.758	.333	.180	.125	.124	.079	+	1.95			1.68		
Non-Adjacent Periods													
1-3	.706	.362	.070	.142	.076	.063	+	.56			.48		
2-4	.707	.343	.085	.131	.065	.056	+	.38			.33		
3-5	.708	.359	.142	.180	.108	.121	-	- .56			- .49		
4-6	.746	.323	.131	.125	.109	.085	+	1.03			.90		
1-4	.655	.264	.070	.131	.057	.043	+	.57			.52		
2-5	.676	.296	.085	.180	.073	.098	-	-1.04			- .93		
3-6	.726	.340	.142	.125	.115	.063	+	2.24	.05		1.94		
1-5	.628	.279	.070	.180	.036	.078	-	-1.72			-1.56		
2-6	.679	.294	.085	.125	.076	.051	+	1.04			.93		
1-6	.634	.252	.070	.125	.069	.024	+	1.83			1.67		
Semester 1,2,3-4,5,6	.788	.535	.109	.155	.112	.098	+	.69			.52		

Table 9 presents the information on the Cross-Lagged Panel Correlations for the all-course classification in which all of the five course division were combined. All but five of the lags are in the positive direction indicating agreement with the absences-cause-poorer-grades prediction. It was felt that a more liberal estimate of the adjusted or "all-test p" levels could be used for the all-course matrix taken by itself. These new levels were computed in the following manner:

$$\frac{1}{100 \times 15 \times 1} \quad \text{or} \quad p = \frac{.00066667}{2} = .00033334 \text{ (two-tailed test)}$$

$$z = 3.40 \text{ for } p \quad .01$$

$$\frac{1}{20 \times 15 \times 1} \quad \text{or} \quad p = \frac{.003333}{2} = .0016667 \text{ (two-tailed test)}$$

$$z = 2.94 \text{ for } p \quad .05$$

Four of the t_r values are significant at the "all-test p" level. One of these four values, the one for fourth and fifth marking periods, is significant in the negative direction indicating a disagreement with the causal prediction. Considering the other three significant t_r values, two of these occur for adjacent marking periods while the remaining value occurs for the nonadjacent third and sixth marking periods. This could be an indication that for the all-course data, the effect of prior absences upon subsequent grades decreases as the length of the lag between marking periods increases. It should be stated that all the results for the all-course data are dubious because of the relatively large N and the presence of more than one class record for many students. An over-all

Table 9

Cross-Lagged Panel Correlations for the All-Course Data (N = 12,3330)

	Test-Retest		Synchronous Cross Variables		Cross-Lagged Correlations		Sign	t_r	Single Test	All Test	t_z'	Single Test	All Test
									p	p		p	p
	Grades	Absences	T_1	T_2	A_1G_2	A_2G_1	$A_1G_2>A_2G_1$						
Adjacent Periods													
1-2	.760	.369	.074	.105	.055	.061	-	- .56			- .47		
2-3	.772	.366	.105	.147	.099	.096	+	.28			.24		
3-4	.776	.340	.147	.114	.126	.073	+	4.89	.001	.01	4.20	.001	.01
4-5	.777	.383	.114	.141	.085	.118	-	-3.12	.001	.05	-2.62	.01	
5-6	.784	.391	.141	.128	.121	.085	+	3.42	.001	.01	2.86	.001	.05
Non-Adjacent Periods													
1-3	.729	.312	.074	.147	.063	.084	-	-1.88			-1.66		
2-4	.740	.321	.105	.114	.084	.059	+	2.26	.05		1.97	.05	
3-5	.745	.341	.147	.141	.109	.098	+	1.01			.87		
4-6	.765	.328	.114	.128	.094	.085	+	.82			.71		
1-4	.700	.241	.074	.114	.060	.050	+	.86			.79		
2-5	.725	.316	.105	.141	.086	.091	-	- .45			- .40		
3-6	.751	.315	.147	.128	.114	.064	+	4.53	.001	.01	3.96	.001	.01
1-5	.677	.286	.074	.141	.046	.072	-	-2.28	.05		-2.05	.05	
2-6	.724	.299	.105	.128	.088	.058	+	2.68	.01		2.37	.05	
1-6	.676	.257	.074	.128	.057	.054	+	.26			.24		
Semester 1,2,3-4,5,6	.821	.528	.126	.147	.123	.100	+	2.42	.05		1.83		

view of the results presented in Tables 4 through 9 indicate that although the magnitude of the synchronous and lagged correlation coefficients were both extremely small, and relatively few significant differences were obtained between the Cross-Lagged Panel Correlations, in the majority of cases for every table except one (Table 6), the direction of the lagging favored the prediction that the predominant direction of causation is from prior absences to subsequent grades.

Analyses and Results for the Multiple Time-Series Design

The results of the Multiple Time-Series Analysis are graphically displayed in Figs. 3 through 8. The all-course graph consists of the pooled results of the previous five individual course categories. The number placed to the left of each curve on every graph indicates in which marking period the "treatment" occurred (i.e., everyone having ten or more absences). The number in parentheses at the right of each curve indicates the number of students whose records were used to compute the values of the points. For the other marking periods the students selected had two or fewer absences. In plotting the control group, it has been matched with treatment group 5. For interpretation, it should be shifted so as to have the same starting point for each of the other groups. However, for economy of graphing, this has not been done.

Many of the groups displayed no particular effect of absenteeism on grades, or, at least, not the predicted drop in the curve for the high absence marking period. There were, however, an encouraging number of groups that did show some indication of a detrimental effect of absenteeism upon grades, especially for the high absence marking period. These groups

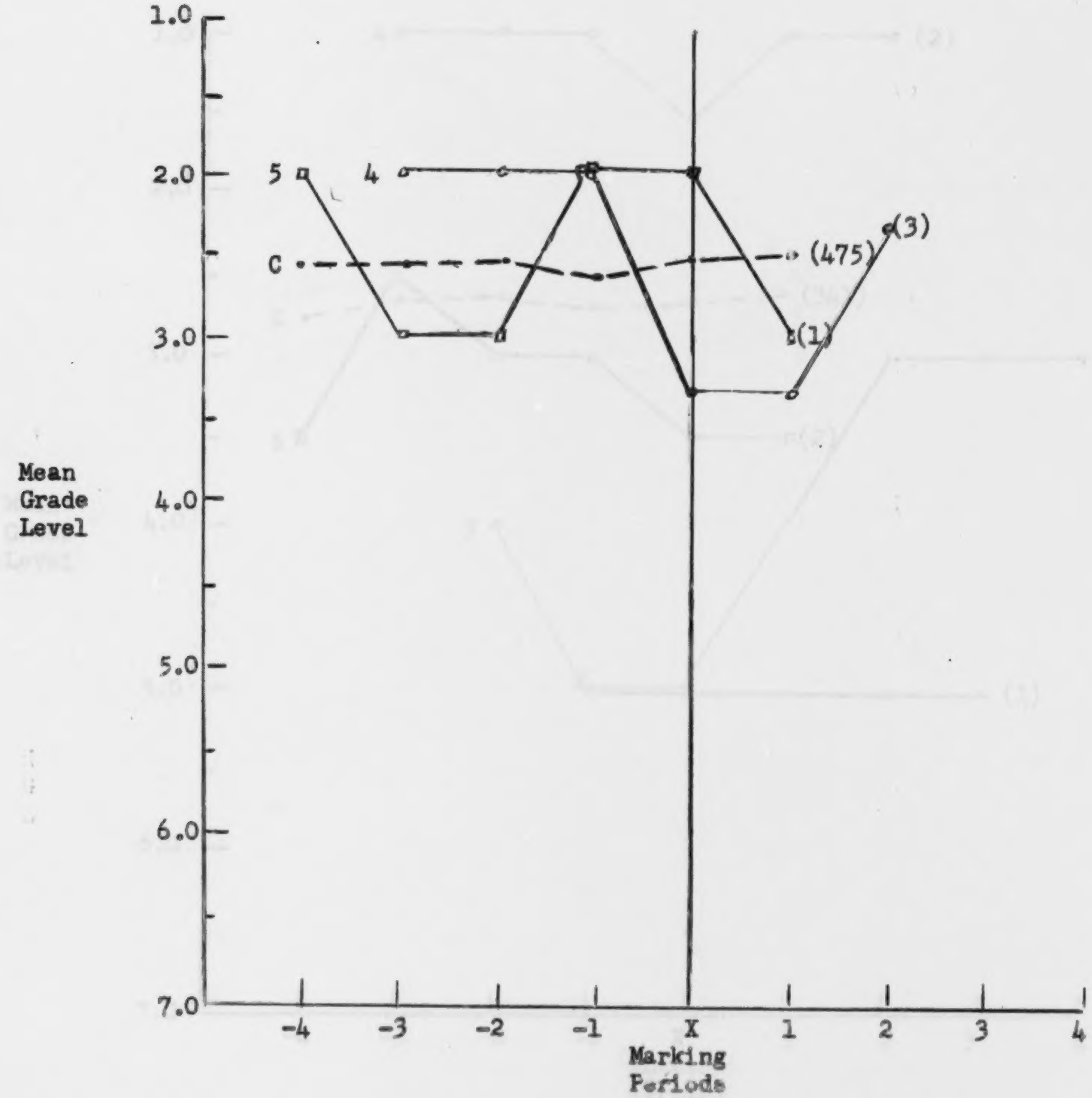


Fig. 3. The effect of absences on grades for first year foreign language courses.

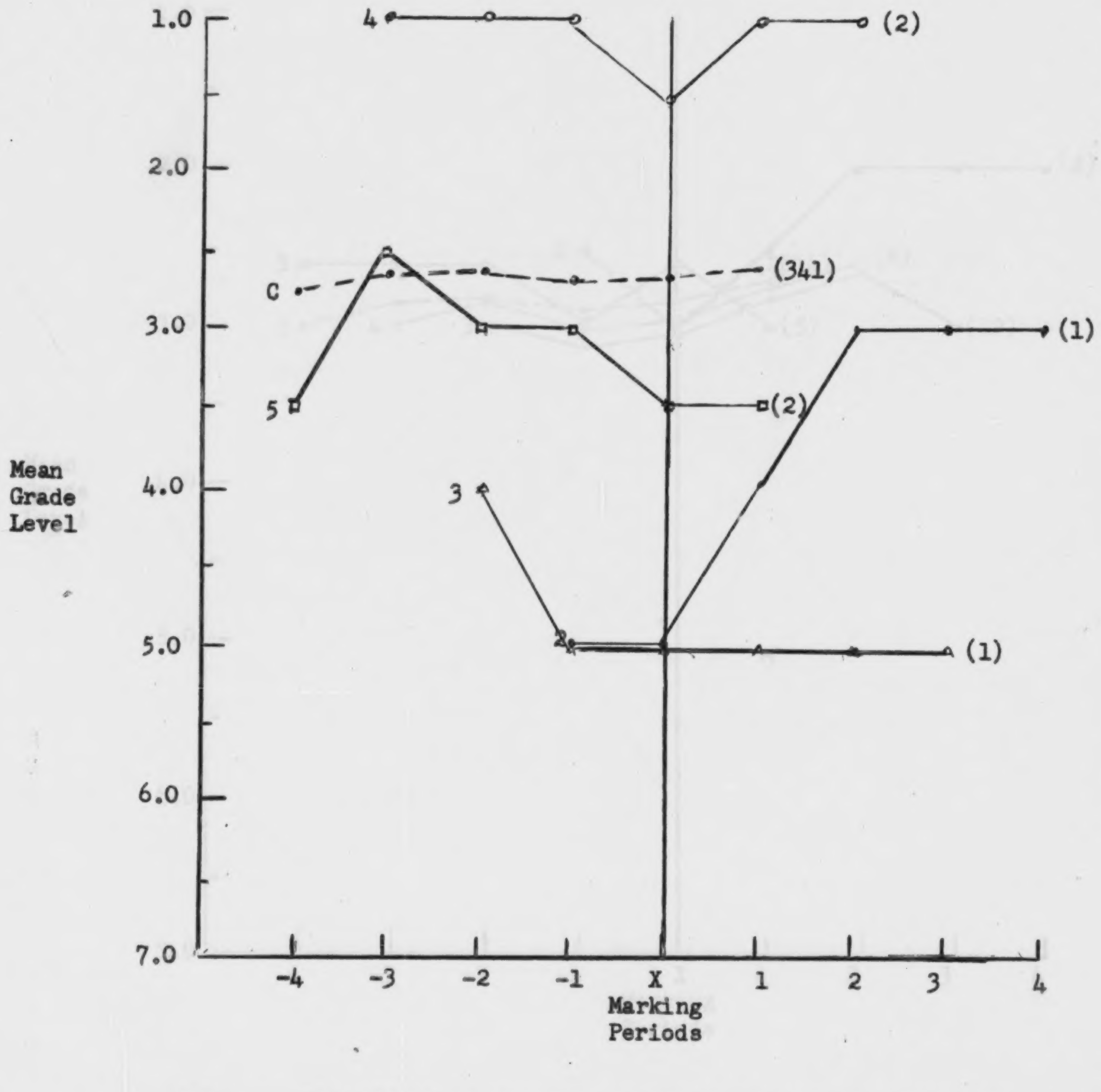


Fig. 4. The effect of absences on grades for second and third year language courses.

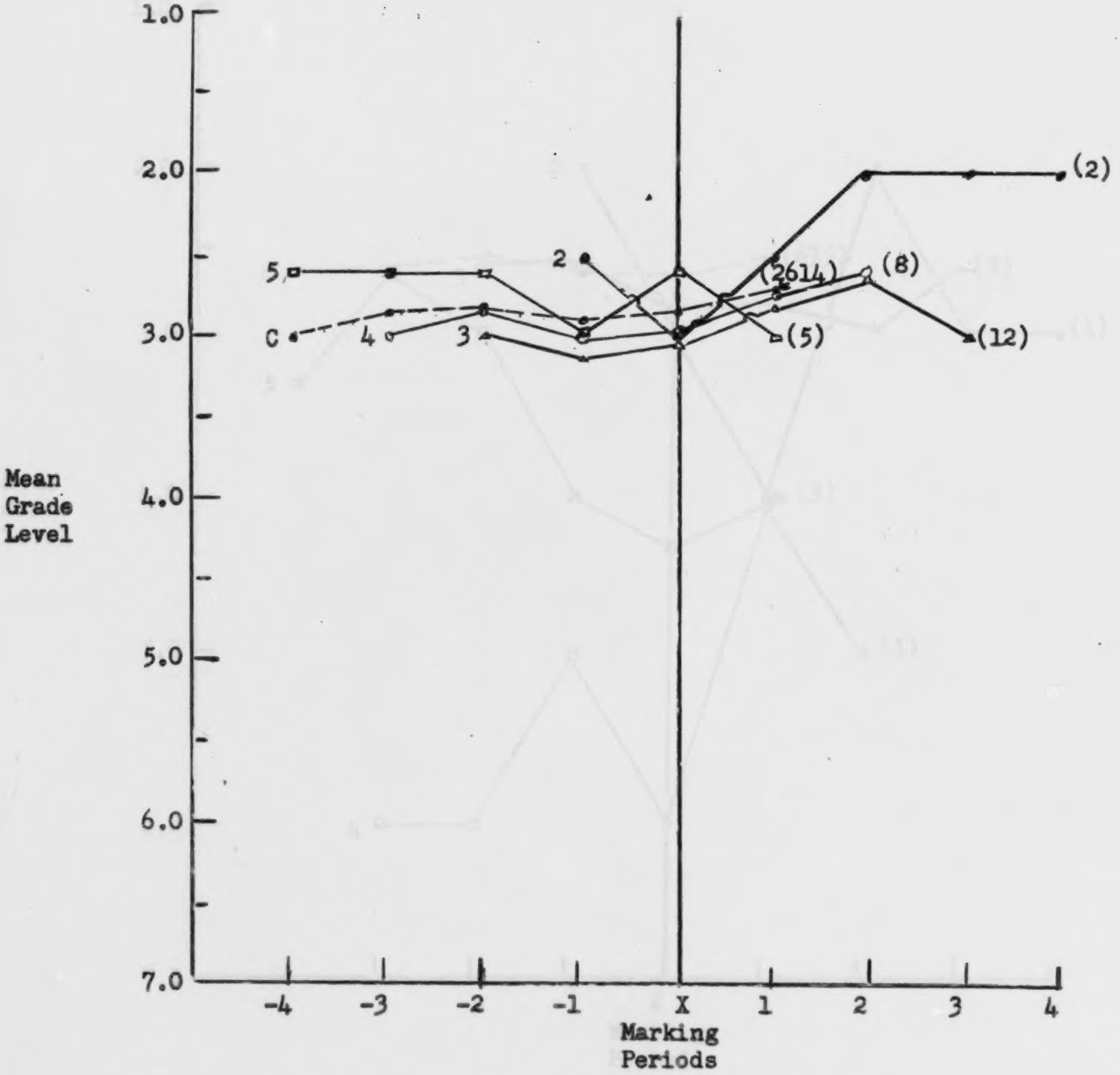


Fig. 5. The effect of absences on grades for combined studies, English, and social studies courses.

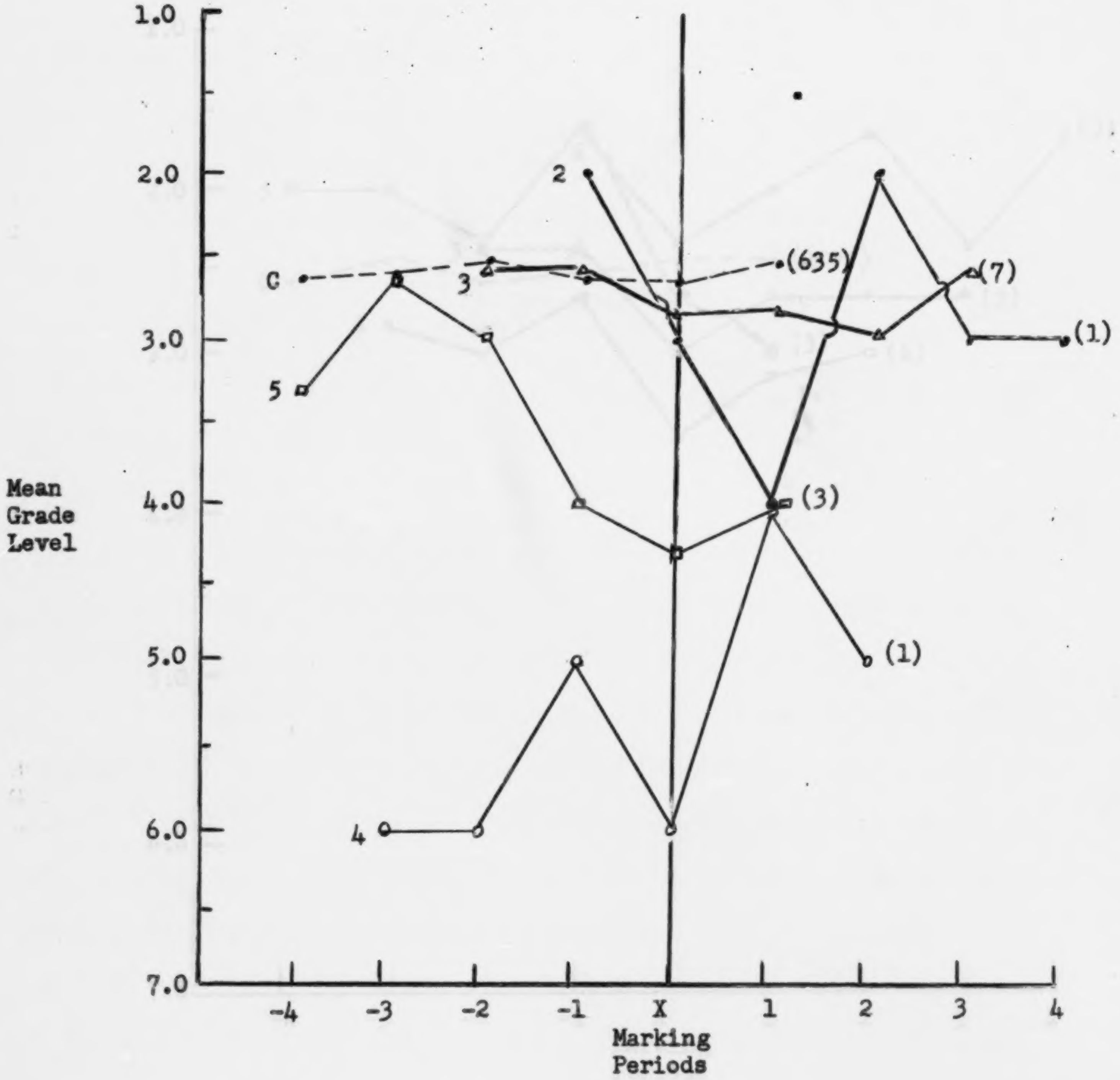


Fig. 6. The effect of absences on grades for science courses.

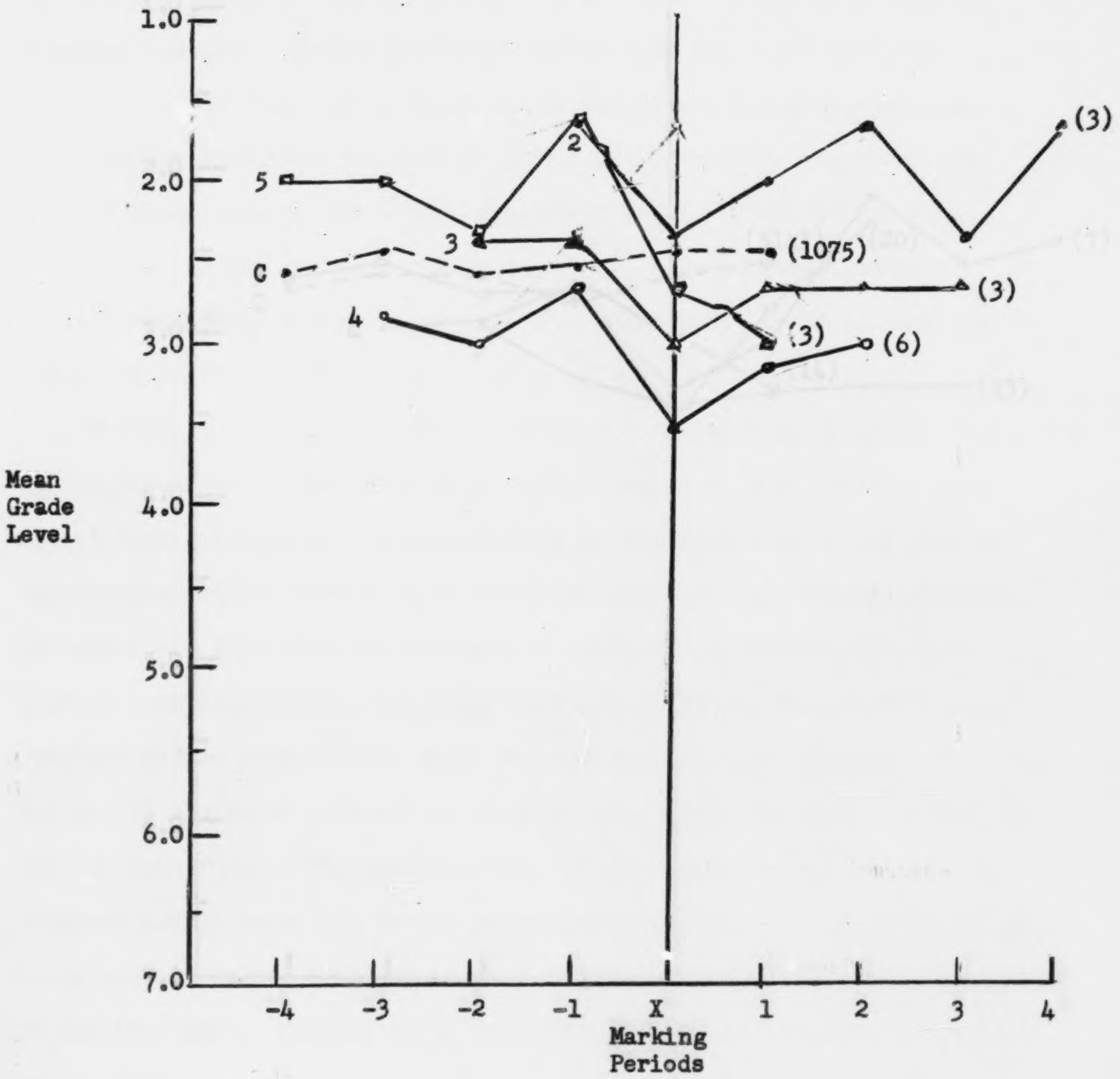


Fig. 7. The effect of absences on grades for mathematics courses.

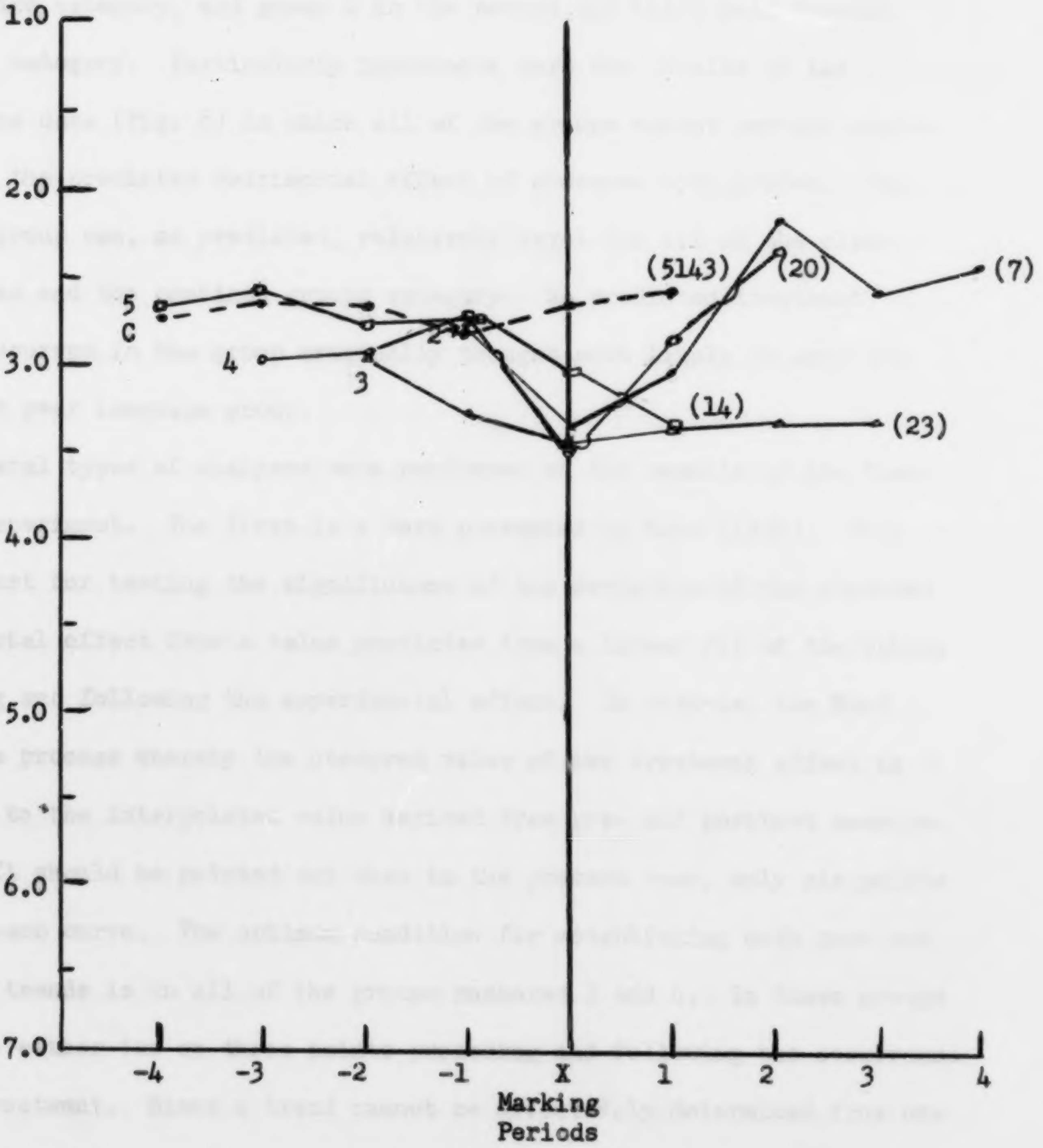


Fig. 8. The effect of absences on grades for all courses combined.

were group 2 in the combined studies category, group 3, 4, and 5 in the mathematics category, and group 4 in the second and third year foreign language category. Particularly impressive were the results of the all-course data (Fig. 8) in which all of the groups except perhaps number 5 showed the predicted detrimental effect of absences upon grades. The control group was, as predicted, relatively level for all of the class-categories and the combined groups category. No predicted treatment effect occurred in the group originally thought most likely to show one: the first year language group.

Several types of analyses were performed on the results of the Time-Series Experiment. The first is a test presented by Mood (1950). This is a t test for testing the significance of the deviation of the observed experimental effect from a value predicted from a linear fit of the values preceding and following the experimental effect. In essence, the Mood test is a process whereby the observed value of the treatment effect is compared to the interpolated value derived from pre- and posttest measurements. It should be pointed out that in the present case, only six points make up each curve. The optimum condition for establishing both pre- and posttest trends is in all of the groups numbered 3 and 4. In these groups there are either two or three points preceding and following the occurrence of the treatment. Since a trend cannot be effectively determined from one point, interpretation of the analysis of groups numbered two and five is difficult because the former has only one pretest point and the latter only one posttest point. In all cases a greater number of points would have been desirable. Nevertheless, it is felt that application of the Mood test to the graph representing the combined groups and the others numbered 3 and 4 mentioned above would furnish some sort of relative numerical reference point from which to interpret the results. Table 10 presents the t's obtained from the Mood test. None of the groups for the combined categories analysis produced highly significant results although groups 3 and 4 approach some relative level of significance. Groups 3 and 4 in the mathematics category were both highly significant. A comparison between the curves for groups 3 and 4 in the mathematics sections with groups 3 and 4 in the combined categories section reveals a high degree of similarity in their shapes. The discontinuities in the curves produced during the high absence period for the mathematics section are more marked than those for the combined categories section. Perhaps the pooling of all the groups produced a leveling tendency which accounts for the lower levels of significance obtained for the combined categories section. However, because the predicted general trend in groups 3 and 4 for the combined categories section prevailed, this does show that for these groups, high absenteeism and relatively lower grades were related. An unusual result was obtained from the application of the Mood test to group 4 in the second and third year languages category: the value of the t was undefined, the reason for this being that the standard error was equal to zero. In effect, a visual inspection reveals that the shape of this curve conformed perfectly to the predictions. However, there were only two people in this group and the actual effect produced in the high absence period was relatively small. The average grade level dropped only one-half of a grade point. It should also be remembered that in all of the cases where verification of the predictions occurred, either statistically or by visual inspection, the N of each group was only a fraction of the entire category

Table 10

The Significance of the Effect of Absence on Grades as Determined by the Mood Test Applied to the Multiple Time-Series Data

Course	Group	N	t	df	p
All-Course Data	2	7	1.296	3	.300
	3	23	2.175	3	.200
	4	20	1.178	3	.200
	5	14	.120	3	.900
Mathematics	3	3	4.914	3	.010
	4	6	3.906	3	.020
Second and Third Year Foreign Languages	4	2	Undefined (Standard Error = 0.0)	3	--

it represented (i.e., the N's used in the Cross-Lagged Panel Correlation analysis). To establish any sort of effect our criterion of absences had to be maintained. A small N reduces the possibilities of specific irrelevancies canceling each other out and increases the probability of plausible rival hypotheses being acceptable explanations for the relationship existing between the two variables. However, whatever plausible rival hypotheses that do exist are not testable in the present methodological analysis and are most likely more individual-specific (attitudes, motivation, etc.) than are grades and absenteeism. Therefore it is felt that the results from the Time-Series Analysis that did correspond to the predictions made do actually show a relationship between grades and absenteeism that is not predominantly explained away by some other unmentioned variable or variables.

The next method of analysis used was the t test between means. In this case, only the combined categories data was involved. The t tests were computed for various points representing different lengths of time before and after the high absence period. The data from all of the groups (except the control) running through each point was pooled so that the t tests determined the degree of significance for the pooled points representing the general trend for the entire combined categories graph. To accomplish this, t tests were only applied to -2 through +2 measurements since it was felt that not enough of the groups simultaneously passed through the points -4, -3, +3, and +4. Again, the assumption that the pooled points represent means of independent random samples is not satisfied. But it was felt that the t test could give useful numerical information about the relative differences between the Time-Series measurements. Table 11 presents the results of the t tests. If the predictions were to be satisfied, no significant differences should have occurred between any of the points except between the marking period just preceding the high absence period (-1) and the high absence period (X), and between the high absence period and the marking period immediately following the high absence period (+1). As Table 11 shows, the predictions were not entirely confirmed. The difference between -1 and X was significant, thus confirming part of the prediction. However, the difference between X and +1 was not significant, perhaps due to a persistence of the effect. Furthermore, the difference between the two posttest measures (+1 and +2), contrary to expectations was significant, whereas the difference between the pretest measurements was not significant as was expected. Also, the difference between the immediately preceding pretest (+1) and the immediately following posttest (-1) was significant.

It is entirely possible that the Time-Series Design can reveal the prolonged effect of high absences in one period upon the subsequent grades of the following working periods. If this was the case, then it could be interpreted that the t-test results revealed this fact. The first significant difference occurred only when the high absence period was introduced. The effect of the high absence period upon the following marking period lowered the grades for that period also which resulted in no significant difference between X and +1. If this were the case, as it was, it would follow that a significant difference would be obtained between the posttest following the high absence period and the pretest immediately preceding it. This also was the obtained result. Examination of the mean grades representing the pooled time measures reveals that the direction of the differences

Table 11

The *t* Test Between Means for the All-Course Data in the Multiple Time-Series Experiment

Points:	Means	N	*t*		
-2	2.913	57	.021	*p*	.450
-1	2.911				
-1	2.881	64	3.321	*p*	.005
X	3.226				
X	3.326	64	1.239	*p*	.150
+1	3.153				
+1	3.105	50	5.071	*p*	.005
+2	2.608				
-1	2.881	64	2.092	*p*	.025
+1	3.153				

was as predicted (i.e., high absences result in lower grades) and substantiates the above interpretation. Remembering that a smaller number represents a higher grade, the mean grade for -1 was 2.881. For X the mean grade was lower, being 3.326. For +1 the grade increased slightly but was still lower than for -1, being 3.153. It should be mentioned that a t test between any two points involved the same groups only. For example, since for -2, group 3, 4, and 5 were involved so that a t test involving -2 and -1 included only those three groups even though an additional group (group 2) ran through -1. Therefore, the mean grade for -1 in the t test for the -1 and -2 is slightly different than the mean grade for -1 in the t test between -1 and X, and -1 and +1. This slight mean change also occurred in the +1 measurement. For both +1 and -1, however, the change was so small that it does not significantly effect the interpretation.

The grades were raised significantly between marking periods +1 and +2. It could be interpreted that the prolonged effect of the high absence period suddenly ceased and grades were restored to their original level as can be seen by examination of the mean grades for -1 (mean grade = 2.881) and +2 (mean grade = 2.608). However, if the only treatment effect on grades was absenteeism, the Time-Series Design should reveal, in an analysis between adjacent points, significant differences only between the measurements adjacent to the presentation of the treatment. So, although two effects could be attributed to absenteeism, i.e., its "onset" (lower grades) and its "offset" (higher grades), other plausible rival hypotheses could be introduced and be rendered acceptable. It cannot be stated with 100 per cent certainty that the same treatment and only the same treatment, i.e., high absences, was involved in both the significant changes observed (between -1 and X, and between +1 and +2). However, it seems reasonable to tentatively conclude that the sudden occurrences of high absences had, at least, an initial detrimental effect upon grades.

The third analysis done on the Time-Series data is a chi-square test presented by Jones and Fiske (1953). This method combines all of the p values obtained from the Mood test and tests the null hypothesis that the composite p value for the several findings could have occurred by chance. This test was employed only on the combined categories data. Again, it is not assured that the various probability values are independent but it was felt that the Jones and Fiske test would provide another general reference point from which to interpret the data. Table 12 presents the results of the Jones and Fiske chi-square test. The null hypothesis cannot be rejected and it must be tentatively concluded that the composite p value for the Time Series findings in the combined categories section could have occurred by chance.

Two of the three analyses employed on the Time-Series data gave indications that high absences are associated with detrimental effect on grades, especially within the marking period that the large number of absences occurred. This result is to a large extent consistent with the correlational data. Examination of the correlation matrices reveals that in many instances the absences and grade records correlate highest within the same grading period. Time-lagged Correlations between absences and grades tend to be lower.

It should be noted that in the Time-Series Experiment the "control" group, i.e., the group that had no high absence marking period, was more

Table 12

The Chi-Square Test for the Composite *p* Value

Obtained from the All-Course Data

$$X^2 = -2 \mathop{}\limits_{i=1}^{k} \ln p_i$$

p	ln	X^2	*df*	Composite *p* Value
.300	-1.20397			
.200	-1.60944	9.05642	8	.50
.200	-1.60944			
.900	-0.10536			
	= -4.52821			

of a comparison group and not, strictly speaking, a "control" group. Nor is the pre- and posttest series of measurements within each particular group a strict control. This design is a longitudinal measurement with shared error of extraneous change agents for adjacent points. There is no reason to believe that at any single time measurement, exactly the same factors are affecting the performance of every group in a cross-sectional analysis. For example, if at time 1 in a particular group A, some extraneous event occurred such that this event was a major determinant in the performance of group A at time 1, then a history confounding has taken place. Now, the experimenter has previously designated time 2 as the treatment time in which group B has received the treatment and group A has not so that group A is supposedly a control group. Furthermore, let us say that group B does show a predicted effect of the treatment and group A does not. Does this conclusively show the existence of only the treatment effect on the basis of a comparison between group A and group B? Not necessarily. It is possible that because of shared error of adjacent points, the extraneous factor that was the causal determinant of group A's behavior at time 1 was also a major factor in the performance of group A at time 2. Then for this unexplained and undetected reason, rather than because of the absence of the treatment effect, did group A remain the same. For the reason that different extraneous factors have affected the performance of group A and group B, a control group-experimental group comparison is not a valid procedure.

Thus, with the Time-Series Experiment, there is never complete experimental isolation so that plausible rival hypotheses cannot be completely ruled out. The experimenter must state, as in the present case, with a certain amount of intuitive certainty that the discontinuity in the series of measurements was due to the hypothesized effect. In the present study, a comparison group was measured separately in five different categories and in the case where all five categories were combined. In every case the comparison group showed no effect in the absence of the treatment (high number of absences) whereas in many cases the treatment groups showed the predicted trends. However, this result may be due to the fact that all of the control groups had a much larger N then the treatment groups which probably balanced out any effects that would have caused the curve to fluctuate if the N had been as small as it was for the treatment groups.

Discussion

Both the Cross-Lagged Panel Correlations and the Time-Series Design should have produced the same results. In only one area, which was not of major interest, was this the case. Both methods of analysis gave evidence that a high number of absences is associated with lower grades within the same marking period. To state this result more generally, it can be concluded that grades and absenteeism are positively correlated within the same time period. However, using the present analyses, unequivocal causal relationships could only have been revealed if the effect of one variable or the other were delayed or lagged. Since the only noticeable impact of absences was on the immediate marking period, the interrupted Time-Series Analysis was not as clear as it would have been if the impact had also been lagged. As the results indicated, some of the Time-Series Analyses tended to show a delayed recovery effect to higher grades after the occurrence of high absence marking period. It would have been at least semirespectable to attribute this trend to the lagged effect of the high absence marking period if the Cross-Lagged Panel Correlations would have also shown a definite causal trend from absence to grades across time periods. As was seen, however, the Cross-Lagged Panel Correlation data showed very little relationship, the nature of which was equivocal, between previous absences and subsequent grades, and vice versa.

The two studies cited previously that resembled, in part, the Cross-Lagged Panel Correlation analysis were those conducted by Odell (1923) and Sandon (1938). Both of these experimenters computed correlations among absences at time 1 and grades at time 2, but not the reverse lagging. Both studies used semester records considerably larger than the six-week marking periods used in the present study. They each correlated first-semester absences with second-semester grades. The results of our Cross-Lagged Panel Correlations involving $r_{A_1G_2}$ agree most closely with Odell's findings. It will be remembered that most of Odell's correlations were smaller than .100 with an over-all correlation of only .08. The time-lagged correlation, using total number of absences computed by Sandon was $r_{A_1G_2} = .26$ which was higher than any obtained in the present Cross-Lagged Panel Correlation analysis. Furthermore, when Sandon correlated the previous number of "spells" of absences with subsequent grades, he obtained an r of .39. He concluded that "frequent short spells of absence are related with educational retardation more than are less frequent longer spells of much greater total duration." In computing the Cross-Lagged Panel Correlations, only the total number of absences were used. If the absence records had been organized into spells of absences, perhaps the correlations in the present investigation would have been higher. However, the sample used in our Cross-Lagged Panel Correlation study was much larger and probably more heterogeneous than that used by Sandon. Sandon's sample consisted of 406 male students whereas this investigator's sample consisted of 12,330 male and female students. With the use of a smaller and more homogeneous sample such as that of Sandon's, there is a greater probability of a plausible rival hypothesis being present. If a dominant and unexplained third factor is operating, it will have a greater chance to be constantly present with a small and homogeneous sample and will therefore result in an "unpure" measure of the variables under investigation. If Sandon had used both sexes and a larger sample, there is the possibility that his correlations would have been considerably decreased. Although Odell used a different age group for his sample (first through eighth grades), his total N equalled 12,773 which is nearly equal to the N of 12,330 used in the present study. Odell's sample also included both males and females.

Neither Odell nor Sandon employed the complete Cross-Lagged Panel Correlation technique and it is impossible to determine whether or not a greater relationship existed between prior grades and subsequent absences. This would have possibly indicated the relative preponderance of cause in their data.

There are many factors in a field study such as this one that could have contributed to the reduction of time-lagged Correlations between absences and grades. Probably one of the most confounding factors is the possible lack of uniformity of grading. It is possible that teachers were more lenient in giving grades to students who were absent from classes due to illness. This would certainly have a leveling effect on the relationship between absences and grades.

As a test of the quasi-experimental methods, the study has been weakened by the very limited, if existent, effect of grades on absences, or by the multiplicity of mutually cancelling effects. One major methodological contribution to the Cross-Lagged Panel Correlation has been made, however. Whereas previous discussions of the method have only considered two competing hypotheses: A increases B, vs. B increases A, in the present setting it became clear that two other ones were also present: A decreases B and B decreases A. Whereas in many settings only two would be at all plausible, in the case of the relationships between absences and grades, all four were. Though for the first year foreign languages the assymetry $r_{A_1G_2} > r_{A_2G_1}$ was persistently found, the recurrent negative correlation between $r_{A_2G_1}$, unexpected by either of the initial competing hypotheses, made the difference unconvincing evidence for the hypothesis that absences-cause-poorer-grades. This finding has both greatly increased the complexity of considerations involved in the Cross-Lagged Panel Correlation, and greatly reduced its apparent utility as a method. While negative correlations where both initial hypotheses predicted positive ones signalled the finding here, such four-way influence would often be covered over by a general level of positive correlation from other sources.

Summary

Two methods of analysis presented by Campbell and Stanley (1963) were applied to the same data for the purpose of testing each method separately and to compare the results of the two methods. The methods were (1) Cross-Lagged Panel Correlation, and the (2) Time-Series Experiment. The data consisted of 12, 330 grade and absence records of students attending the Evanston Township High School, Evanston, Illinois. This data represented grades and absences for six marking periods, each period lasting approximately six weeks.

The Cross-Lagged Panel Correlations reveal time-lagged relationships between variables. Thus, absence records for one marking period were correlated with grade records for every other marking period. All combinations of time-lagged Correlations among grades and absenteeism were obtained. It was expected that absences at time 1 would have a detrimental effect upon grades at time 2 and that the magnitude of this relationship would be greater than that between grades at time 1 and absences at time 2. In other words, the predominant causal direction was expected to be from prior absences to subsequent grades, and not from prior grades to subsequent absences. The Time-Series Experiment consisted of plotting the mean grades for each marking period. Data was selected in such a way so that a minimum of absences occurred in every marking period except one. In this single marking period a maximum number of absences occurred. It was expected that the average grades for the maximum absence marking periods would be significantly lower than those for the minimum marking periods. This result would further substantiate a causal effect from absences to grades.

The results obtained from the Cross-Lagged Panel Correlation analysis were not conclusive in confirming or refuting the predictions made. Only a very small time-lagged relationship among number of absences and grades was revealed. The predominant causal direction between the two variables was not revealed, i.e., correlations between absences at time 1 and grades at time 2 were not consistently or for the most part, significantly larger than correlations between grades at time 1 and absences at time 2.

The Time-Series Analysis revealed more positive results. In many of the curves plotted, there occurred a noticeable drop during the maximum absence marking period indicating lower grades. In a few instances this change in intercept was significantly different from an interpolated intercept value. It must be stated, however, that there were more Time-Series curves which did not reveal the predicted effect than there were curves that did. Also, most of the curves that tended to confirm the predictions by a visual analysis did not do so by the statistical analyses performed. In conclusion, the prediction that the results of the Cross-Lagged Panel Correlation and Time-Series Analysis would be similar was not confirmed. The Cross-Lagged Panel Correlations were not definitive in revealing the direction of causation between grades and absences. The Time-Series Analysis revealed some detrimental effect of absences upon grades, but the effects that did occur were small.

A major contribution to the understanding of the Cross-Lagged Panel Correlation was made when it became clear that four causal hypotheses are competing in the analysis, not two. Pairs of these four are confounded, reducing the interpretability of a significant assymetry between the two Cross-Lagged values, at least for those settings in which all four have some plausibility.

References

Amthauer, R. Ergebnisse einer Studie über krankheitsbedingte Fehlzeiten. [Results of a study on absenteeism caused by sickness.] *Psychologische Rundschau*, 1963, 14, 1-12.

Anikeef, A. M. The relationship between class absences and college grades. *Journal of Educational Psychology*, 1954, 45, 244-249.

Butler, C. H. School achievement and attendance. *School Review*, 1925, 33, 450-452.

Campbell, D. T. From description to experimentation: interpreting trends as quasi-experiments. In C. W. Harris (Ed.) *Problems in Measuring Change*. Madison, Wis.: Univer. Wisconsin Press, 1963. Pp. 212-242.

Campbell, D. T., Miller, N., Lubetsky, J., & O'Connell, E. J. Varieties of projection in trait attribution. *Psychological Monographs*, in press.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.) *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963. Pp. 171-246.

Caro, P. W., Jr. The effect of class attendance and "time structured" content on achievement in general psychology. *Journal of Educational Psychology*, 1962, 53, 76-80.

Crider, B. The effect of absence on scholarship. *School and Society*, 1929, 30, 27-28.

Cutler, Vera. The relation between attendance and achievement in the third grade of the San Jose School. Master's Thesis, Univer. of Mexico, 1936.

Edmiston, R. W., Hinton, M. E., & Rasor, F. Special emphasis to improve attendance. *Journal of Educational Research*, 1947, 41, 35-40.

Edwards, A. L. *Experimental Design in Psychological Research*. (Rev. ed.) New York: Rinehart, 1960.

Feingold, G. A. Scholarship, attendance,and deportment. *School and Society*, 1928, 28, 333-335.

Finch, F. H., & Nemzek, C. L. Attendance and achievement in high school. *School and Society*, 1935, 41, 207-208.

Finch, F. H., & Nemzek, C. L. Attendance and achievement in secondary school. *Journal of Educational Research*, 1940, 34, 119-126.

Fisher, R. A., & Yates, F. *Statistical Tables for Biological, Agricultural and Medical Research*. (4th ed.) New York: Hafner, 1953.

Heilman, J. D. The relative effect upon educational achievement of some hereditary and environmental factors. *The Twenty-Seventh Yearbook of the National Society for the Study of Education*, 1928, 2, 35-65.

Hersov, L. A. Persistent non-attendance at school. *Journal of Child Psychology and Psychiatry*, 1960, 1, 130-136.

Jensen, M. B. The influence of class size upon pupil accomplishment in high school algebra. *Journal of Educational Research*, 1930, 21, 337-356.

Jones, L. Class attendance and college marks. *School and Society*, 1931, 33, 444-446.

Jones, L. V., & Fiske, D. W. Models for testing the significance of combined results. *Psychological Bulletin*, 1953, 50, 375-382.

Monroe, W. S. (Ed.) *Encyclopedia of Educational Research*. (1st ed.) New York: Macmillan, 1941.

Monroe, W. S. (Ed.) *Encyclopedia of Educational Research.* (2nd ed.) New York: Macmillan, 1950.

Mood, S. M. *Introduction to the Theory of Statistics.* New York: McGraw-Hill, 1950.

Nelson, M. J. An experiment with optional attendance. *School and Society,* 1937, 45, 414-416.

Nemzek, C. L. The value of amount of tardiness and absence for direct and differential prediction of academic success. *Journal of experimental Education,* 1938, 7, 4-10.

Odell, C. W. The effect of attendance upon school achievement. *Journal of Educational Research,* 1923, 8, 422-432.

Pearson, K., & Filon, L.N.G. Mathematical contributions to the theory of evolution. Trans. Roy. Soc. (London), Series A, Vol. 191. Pp. 259-262. Cited by C. C. Peters & W. R. Van Voorhis in *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill, 1940, p. 185.

Peters, C. C., & Van Voorhis, W. R. *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill, 1940.

Platt, S. W. Intrinsic causes for public school absences. *School and Society,* 1943, 57, 307-308.

Ryan, T. A. Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin,* 1960, 57, 318-328.

Sandon, F. The comparative effect of school progress of many short and one long absence in secondary school. *British Journal of Psychology,* 1938, 8, 172-177.

Stephens, J. M. The Influence of School on the Individual. Ann Arbor Michigan: Edwards Brothers, 1933.

Thistlethwaite, D. L., & Campbell, D. T. Regression-discontinuity analysis: an alternative to the ex post facto experiment. Journal of Educational Psychology, 1960, 51, 309-317.

Tyerman, M. J. A research into truancy. British Journal of Educational Psychology, 1958, 28, 217-225.

Wetzel, H. A. School attendance and scholarship. School Review, 1928, 36, 118-120.

Woolf, Henriette K. The relative influence of some environmental factors and mental age on educational age. Unpublished Master's Thesis, Colorado State Teacher's College, Greeley, Colorado, 1926. Cited in J. D. Heilman. The relative effect upon educational achievement of some hereditary and environmental factors. The Twenty-Seventh Yearbook of the National Society for the Study of Education, 1928, 2, 35-65.

Young, A. J. Truancy: a study of mental, scholastic, and social conditions in the problem of non-attendance at school. British Journal of Educational Psychology, 1947, 17, 50-51.

Causal Relations in Attitude Change as Demonstrated Through the Cross-Lagged Panel Correlation*

Richard M. Rozelle

Northwestern University

---

*Supported by Project C-998, Contract 3-20-001, Education Media Branch, U.S. Office of Education, Department of Health Education and Welfare, under Title VII of the National Defense Education Act.

This report presents the results of a quasi-experimental retrospective analysis of two sets of attitudinal data. One set of data was collected at Vassar College (Webster, Freedman, & Heist, 1961; Bereiter, 1963) while the other set consisted of measures taken from small, isolated research groups stationed in the Antarctic (Gunderson and Nelson, 1962). In both sets of data, multiple measurements of attitudes were taken with a substantial time period between the measurements.

The analysis used on the data is the Cross Lagged Panel Correlation, first presented by Campbell and Stanley (1963) who have presented a number of quasi-experimental designs. The program of quasi-experimental analysis seeks to extend into the natural field situations the orientation of experimental analysis in the delineation of cause-and-effect relationships. Studies are classified as quasi-experimental where random assignment to treatments is not possible.

Cross Lagged Panel Correlations are correlations between measures of variables taken at different times for the purpose of comparing the relative magnitude of the antecedent-subsequent relationships that exist. The principal reason for using the CLPC, is to aid in deciding upon the direction of the causation in the situation where two series of data correlate. If C is cause and E is effect, then $r_{C_nE_{n+1}} > r_{C_{n+1}E_n}$ with n designating the relative time the measurement is taken (Campbell, 1963, p. 236). For example, in the Gunderson and Nelson data, two of the attitudinal variables that were measured at two different times were Group Achievement (GA) and Equalitarian Atmosphere (EA). The question is then asked; given that these two variables are correlated, does prior GA lead to greater subsequent EA or does prior EA lead to subsequent greater GA? In the notation presented above, the question is: $r_{GA_1EA_2} > r_{GA_2EA_1}$ or $r_{GA_1EA_2} > r_{GA_2EA_1}$? These time-lagged correlations between the two variables should reveal an index of relative preponderance and therefore the direction in which a causal inference may be made.

A Montecarlo Examination of Tests of Significance

One of the major problems encountered when using correlation coefficients, especially the time-lagged variety, is testing the significance of their differences (Rozelle, 1965). When the two coefficients are computed on the same population, and the arrays of the coefficients involved are correlated--even if none are repeated--then this correlation of arrays reduces the sampling variation and should be taken into account in the error term. Peters and Van Voorhis (1940, p. 185) provide a Pearson and Filon test for this situation, as well as the complex procedure for computing the correlation between correlation coefficients used to reduce the error term. This involves all of the other correlations among the four variables presented in the first four columns of Table 1. The resulting t statistic is designated as t_p. More orthodox today is the use of Fisher's z' transformation of correlation coefficients in factors of their significance. No z' formula reducing the error term for the correlation between correlation coefficients is available. These run consistently slightly smaller than do the other t's, due to the larger error term. The z' values are designated as $t_{z'}$ in Table 1.

Insert Table 1 about here

To examine the appropriateness of these tests of significance, they have been tried out in a "montecarlo" study, that is, on samples of correlated variables computer-generated from combinations of normal random numbers. In all cases the null hypothesis was "true," that is, in terms of the sampling model involved, the two hypothetical cross-lagged correlations were equal, any differences being due to sampling error. The values were taken from correlation matrices computed previously for another project. Successive sets of 4 adjacent rows and columns were arbitrarily labeled A_1 (variable A at time 1), A_2, B_1, and B_2. There were 364 such sets, varying in N from 20 to 100, and in "true" correlation from .25 to .90. By the Pearson and Filon test, there were 13 ts significant at the $p < .05$ level where chance expectancy would be for 18. There were but 2 values significant at the $p < .01$ level, where chance expectancy would be 3.6. By t_z, none reached significance at the $p < .01$ level, and only 7 of the expected 18 at the $p < .05$ level. Thus the z' test is much too conservative, the Pearson and Filon test much more accurate, although slightly conservative, too.

It is not unusual that when comparing correlated variables, the size of the correlation matrix is larger than the simple 2 x 2 variety. In the Rozelle study for example, each correlation matrix involved 16 variables. In addition, five of these 16 x 16 matrices were computed, increasing the number of correlational comparisons to a considerable degree. The sample data considered in the montecarlo study consists of 364 tests. Were only one t value to be examined for significance, or were all t values uniformly high, the standard single-valued criteria for significance levels obtained directly from any t-table would seem appropriate. Where, however, many tests are computed and only a few reach such significance levels, even these few may be suspect of occurrences made more probable by the many tests involved. While no entirely satisfactory solution to this problem is known, Campbell, Miller, Lubetsky, & O'Connell (1965) have proposed a number of baselines for comparison in this multiple-comparison situation.

Another baseline for comparison proposed by Campbell, Miller, Lubetsky, and O'Connell, is a more conservative revision of the significance level. This may be computed by asking how often would a value of this magnitude occur by chance in 100 replications of the N tests. The $p < .01$ level becomes the $\frac{1}{100 \times N}$ level in the normal curve table. For the $p < .05$ level, the new value is $\frac{1}{20 \times N}$. With the z' values obtained from Fisher and Yates (1953, p. 59), Table $VIII_4$, the adjusted significance levels are:

(a)

$\frac{1}{364 \times 100}$ or $p = .00002747$, or for a two tailed test, .00001374

for which $z' = 4.19$ for $p < .01$

(b)

$\frac{1}{364 \times 20}$ or $p = .00013736$, or for a two-tailed test, .00006868,

for which $z' = 3.81$ for $p < .05$

There adjusted or whole set significance levels are indicated in Table 1 as "all-test p" values. Examination of the sample data reveals that no t value, even using the larger Pearson and Filon value, is significant at either the .01 or .05 level. The largest t_r is but 2.87. This is, of course, as it should be. Since the t's are but chance variations on a true difference of zero, and one would expect by chance alone only one t of 3.81 in 20 replications of the sample of 364, and a t of 4.19 only once in 100 replications of the 364.

Ryan (1960) has distinguished between error rate per experiment and error rate experimentwise. Error rate per experiment is defined as the number of comparisons incorrectly called significant, divided by the total number of experiments. This error rate is tested by the "all-test p" values in Table 1. Error rate experimentwise is defined as the number of experiments containing erroneous statements of significance, divided by the total number of experiments. Ryan states that the error rate experimentwise can never be greater than the error rate per experiment because there cannot be more experiments with erroneous conclusions than there are errors in individual comparisons. The error rate experimentwise is computed by the following formula:

$$1 - (1 - x)^k = .05 \quad (.05 \text{ level})$$
$$1 - (1 - x)^k = .01 \quad (.01 \text{ level})$$

For both levels, $k = 386$ or the total number of time-lagged comparisons recorded. The z values obtained from this method were $z' = 4.18$ for $p < .01$ and $z' = 3.80$ for $p < .05$. Those values are almost identical to and slightly smaller than those obtained from the error rate per experiment method. For purposes of this study, then, the distinction between error rate per experiment and error rate experimentwise is not important in that it has no differentiating effect upon the statistical interpretations of the comparisons made. It can be noted that Roselle (1965), with a $k = 75$, discovered the same lack of difference between significance levels of the error rate per experiment technique and those of the error rate experimentwise technique.

Vassar Students Measured as Freshmen and Seniors

Data collected at Vassar College (Webster, Freedman, & Heist, 1961) offered an opportunity to utilize the CLP Correlational technique as a means of analysis to test several causal hypotheses.

The central hypotheses concern the variables of Ethnocentrism (E) and

Authoritarianism or Fascism (F). Adorno, et al. (1950) constructed attitude scales to measure E and F. The E-scale was initially constructed and contained items of relative specific content, mentioning actual minority groups by name. The E scale did not conceal its aim: to measure attitudes of prejudice. After a number of investigations had been done with the E scale which showed high correlations with variables dealing with political, economic, and social matters, it was felt that perhaps the E scale would correlate with more basic, underlying personality variables. If a more general attitude scale could be constructed, i.e., one which would measure prejudice without appearing to have this aim and without mentioning the name of any minority group and that would yield a measure of antidemocratic attitudes at the personality level, and the E-scale would correlate highly with this new scale, then ". . . it could be taken as evidence that anti-Semitism and ethnocentrism were not isolated or specific or entirely superficial attitudes but expressions of persistent tendencies in the person" (Adorno, et al., 1950, p. 223). The new, more general attitude scale, the F scale, was constructed, and was shown to be an accurate predictor of scores obtained on the E scale. The F scale is considered to be a more general and direct measure of the all-encompassing personality variable, authoritarianism. The F scale was shown to be correlated with a wide range of social and personality dispositions, only one of which was ethnocentrism. Thus, in this context, ethnocentrism measured by the E scale, is considered a more specific variable, a symptom of the more general personality variable labeled authoritarianism which is measured by the F scale. If this be the case, then it seems reasonable to hypothesize that if there is a causal relationship between ethnocentrism and authoritarianism, the greatest direction of cause should be from F to E, not from E to F. In other words, an individual possessing a certain degree of authoritarianism in their personality, should develop a corresponding level of ethnocentrism in their attitudes, rather than vice versa. The tone of specific attitudes is determined by the general outlook a person possesses. Thus, if the E and F scales were administered simultaneously at two different points in time, the prediction, using CLPCs would be

$${}^{r}F_{1}E_{2} > {}^{r}F_{2}E_{1}$$

An alternative hypothesis regarding authoritarianism and ethnocentrism is suggested by data obtained from an experiment conducted by Hovland, Lumsdaine, & Sheffield (1949). In this well-known experiment conducted during World War II, American soldiers were shown a film, "The Battle of Britain," the purpose of which was to investigate the effects of this dramatic, anti - Nazi film upon the attitudes of the service men. Attitude measurements were taken on some groups five days after the film presentation and on other groups, nine weeks after the presentation. In general, the findings indicated that immediately the film was shown, attitudes that were affected most were those dealing with specific aspects of the British war effort such as the competence of the RAF pilots, purpose of Nazi Bombing of Britain, etc. However, groups who were administered the opinion questionnaire nine weeks after the film was presented, showed greater attitude change in areas dealing with general attitudes toward the British such as "British are doing their fair share of the fighting." From this data confirmed the hypothesis that changes in opinions of a general rather than a specific nature may show increasing effects with a lapse of time. This could be due to the fact that attitudes are initially learned in connection with specific information and, with time, generalize to a broader scope. If this is true, then a specific factor such as ethnocentrism would be a more determining factor upon subsequent authoritarianism than vice versa, i.e., if a causal relationship exists between E and F, then, according to the Hovland _et al_. data, the greater direction of cause would be from E to F, not from F to E. In CLPC language, the prediction is:

$$r_{F_2E_1} > r_{F_1E_2}$$

Thus, we have two conflicting predictions:

(a) according to Adorno, _et al_.

$$r_{F_1E_2} > r_{F_2E_1}$$

(b) according to Hovland _et al_.

$$r_{F_1E_2} < r_{F_2E_1}$$

The data collected at Vassar College consisted, in part, of the same attitude and personality questionnaires administered to three different classes at the beginning of their freshmen year and again at the end of their senior year. The questionnaires contained a number of variables designed to measure "the expression of impulses and certain aspects of ego functioning and development in order to obtain optimum descriptions of women students." Among these variables were the E scale and the F scale. The graduating classes of 1957, 1958, and 1959 were used as subjects in the study. From this data we have computed a correlational matrix which includes scores of E and F, as well as seven other variables, correlated over time, i.e., the freshmen and senior scores have been correlated. Table 2 shows the correlations for the class of 1957, Table 3 for 1958, and Table 4 for

1959. Table 5 presents the pooled correlations for all three classes combined.

The adjusted error rate per experiment or "all test p" level is applied to Tables 2 through 4.

There are 3 matrices with 36 CLPC comparisons per matrix.

$$\frac{1}{36 \times 3 \times 100} \text{ or } p = \frac{.00009259}{2} = .000046295 \text{ (two tailed test)}$$

$$z' = 3.91 \text{ for } p < .01$$

$$\frac{1}{36 \times 3 \times 20} \text{ or } p = \frac{.004621}{2} = .0023105 \text{ (two tailed test)}$$

$$z' = 2.50 \text{ for } p < .05$$

It is felt that on the basis of past experience, the estimate of error rate experimentwise is of no significant difference from the "all-test" or error rate per experiment criterion. Thus computation of error rate experimentwise estimate is omitted from all subsequent analyses. Table 2 reveals that $r_{F_1E_2} > r_{F_2E_1}$, indicating that the greatest relative direction of causation is from prior authoritarianism to subsequent ethnocentrism. This would be in agreement with the Adorno _et al_. hypothesis. However, the difference between these two CLPCs is not significant, either at the single test level or the more conservative all-test _p_ level for both the .01 and .05 levels.

Table 3 shows that for the class of 1958, $r_{F_1E_2} = r_{F_2E_1}$. This result gives no indication of the preponderance of causal direction and would only give evidence of balanced, bidirectional cause-effect relationship between the two variables. Table 4 presents results quite similar to those of Table 3 except that $r_{F_1E_2}$ is slightly, but not significantly, smaller than $r_{F_2E_1}$, which would, at least, be in the direction of agreeing with the Hovland _et al_. hypothesis. Table 5 reveals that with all of the data pooled, $r_{F_1E_2} > r_{F_2E_1}$

which is in the direction of confirming the Adorno authoritarian personality hypothesis. It was felt that more liberal estimate of the adjusted or "all-test *p*" levels could be used for the single matrix representing all of the data combined. These new levels were computed in the following manner:

$$\frac{1}{36 \times 1 \times 100} \text{ or } p = \frac{.00027777}{2} = .00013888 \text{ (two tailed test)}$$

$$z' = 3.64 \text{ for } p < .01$$

$$\frac{1}{36 \times 1 \times 20} \text{ or } p = \frac{.00138888}{2} = .0006944 \text{ (two tailed test)}$$

$$z' = 3.20 \text{ for } p < .05$$

However, as in the other three cases, the difference between these two CLPCs is not significant at any level. An overall view of the F-E CLPCs would seem to indicate evidence favoring the Adorno *et al*. hypothesis, i.e., there is greater direction of cause from the more general authoritarian variable to the more specific ethnocentrism variable than from ethnocentrism to authoritarianism. Due to lack of any significance in differences, however, it is difficult to meaningfully or conclusively interpret the results.

The other variables presented in Tables 2 through 5 are described by Webster, Freedman, & Heist (1961) in the following manner:

SM = Social Maturity. Low scorers are authoritarian, compulsive, rigid, punitive, submissive to power, conventional, cynical, anti-intellectual, and emotionally suppressed. Hi-scorers are relatively free of these characteristics. The scale provided a measure of F that was less ideological than the original F scale.

IE = Impulsive Expression. High scorers, in contrast to low scorers, have a greater readiness to express impulses, or to seek gratification in them in overt action or in conscious feeling and attitude. High scorers are usually dominant, aggressive, autonomous, exhibitionistic, and express interest in sex, excitement, and change.

DS = Developmental Status. A scale made up of attitude items that distinguish younger from older students. In a sense such items reflect developmental from the freshman to senior year. High scorers (seniors) in comparison with low scorers (freshmen) are flexible and uncompulsive, impunitive toward persons but critical of the institutional authority of family, state, or religion. They are also intraceptive, nonconforming, free of cynicism, realistic, and mature in interests. Content from the two previous scales also appears in this scale.

DC = Dominance and Confidence. High scorers endorse items avowing self confidence, imperturbability, and ability to lead.

RS = Repression and Suppression. High scorers will tend, often quite unconsciously, to reject (answer "false") a disproportionately large number of statements which seem incorrect, unconventional because he has a strong need to present himself in a "favorable" light. Low scorers are opposite. They show imprudence, unconventionality, and lack of inhibition. The "realistic" place to score on this scale is near the mean.

SI = Social Integration. High scorers are likely to be conventional and free of symptoms of social alienation; low scorers usually feel that they are unhappy social isolates.

MR = Masculine Role. High scorers are usually active persons with interests and attitudes characteristic of men in our culture; low scorers are passive, acquiescent, and conventionally feminine.

Although these variables are of no theoretical interest in the present report, significant differences among them are reported now.

In Table 2, for the class of 1957, there are two significant differences between CLPCs at the all-test p criterion. The difference between DS and F is significant at the $p < .01$ level for both t_r and $t_{s'}$. Both CLPCs are negative ones with $DS_1F_2 < DS_2F_1$, thus a greater F characteristic causes less subsequent DS, or, as the case may be, less prior F results in more subsequent DS. The other significant difference in Table 2 is between SM and F, and is significant for the t_r at $p < .01$ and $t_{s'}$ at $p < .05$. These also are negative CLPCs and indicate that $SM_1F_2 < SM_2F_1$. Other differences in Table 2 are significant only at the standard single-test p level. With significance levels given for t_r only these are, $SI_1DC_2 > SI_2DC_1$ ($p < .05$), $SI_1DS_2 < SI_2DS_1$ ($p < .05$), $RS_1SM_2 > RS_2SM_1$ ($p < .05$), $DS_1SM_2 < DS_2SM_1$ ($p < .05$),

$SM_1E_2 < SM_2E_1$ (both negative, $p < .05$). An important problem is encountered with the significant difference between $SI_1SM_2 > SI_2SM_1$ in that $r_{SI_1SM_2}$ is positive and $r_{SI_2SM_1}$ is negative. Thus, because in Tables 3 and 4 $r_{SI_2SM_1}$ is positive, prior high SM can result in either a high or low degree of subsequent SI and prior low SM can result in either a low or high degree of SI. Thus four causal hypotheses are competing in the analysis, not two. Pairs of these four are confounded, reducing the interpretability of a significant assymetry between the two Cross-Lagged values. This problem was also encountered by Rozelle (1965) and is discussed more fully in that report.

For the class of 1958 (Table 3) there are no significant differences between any of the CLPCs at the all-test *p* level. Two significant t_r's at the $p < .05$ level are reported for the single test *p* criterion and are: $SI_1IE_2 < SI_2IE_1$ (both negative) and $RS_1IE_2 < RS_2IE_1$ (both negative).

The class of 1959, as was the case with that of 1958, produced no significant differences among any of the CLPCs for the all-test *p* criterion. For the single-test *p* criterion with probability levels given only for t_r, the following CLPC comparisons were significant: $SI_1DC_2 > SI_2DC_1$ $p < .01$, $DC_1RS_2 < DC_2RS_1$ $p < .05$, $DC_1IE_2 < DC_2IE_1$ $p < .05$ (both negative), $MR_1SM_2 < MR_2SM_1$ $p < .01$, $RS_1IE_2 < RS_2IF$ $p < .05$ (both negative), $DS_1SM_2 < DS_2SM_1$ $p < .01$, $SM_1IE_2 < SM_2IE_1$ $p < .01$.

The combined groups matrix produced three CLPC comparisons that were significant at the all-test *p* criterion especially computed for these pooled data correlations: $DS_1SM_2 < DS_2SM_1$ $p < .01$, $DS_1F_2 < DS_2F_1$ $p < .01$, (both negative), and $SM_1IE_2 > SM_2IE_1$ $p < .05$. Other differences that were only significant at the single-test *p* level with probability levels given for t_r only are: $SI_1DC_2 > SI_1DC_2$ $p < .05$, $SI_1DS_2 < SI_2DS_1$ $p < .01$ (both negative),

$SI_1SM_2 < SI_2SM_1$ $p < .01$, $SI_1IE_2 < SI_2IE_1$ $p < .05$ (both negative), $SI_1F_2 < SI_2F_1$ $p < .05$, $SI_1E_2 < SI_2E_1$ $p < .01$, $DC_1RS_2 < DC_2RS_1$ $p < .05$, $MR_1RS_2 < MR_2RS_1$ $p < .05$, $MR_1SM_2 < MR_2SM_1$ $p < .05$, $RS_1DS_2 > RS_2DS_1$ $p < .05$ (both negative), $RS_1SM_2 > RS_2SM_1$ $p < .01$, $RS_1IE_2 < RS_2IE_1$ $p < .05$ (both negative), $SM_1F_2 < SM_2F_1$ $p < .01$ (both negative), $IE_1F_2 < IE_2F_1$ $p < .05$ (both negative).

It is interesting to point out that in Tables 2 through 5, in every case where a significant difference occurred in which F or E and any one of the other seven variables were involved, either at the all-test p or single test p criteria, the CLPC that was largest was that which contained F_1 or E_1 indicating a greater causal relationship from a prior F or E to a subsequent effect on the other variable. Whether this finding is relevant to the general factor-causes-specific factor hypothesis discussed with the F and E scales is not known because, although it would seem that the F scale is a more general measurement than the other seven variables, the other seven scales have not been examined to determine their relationship to the E scale. Although in many the absolute value of the correlation here is of such small magnitude that their meaning is difficult to assess, perhaps further investigations would reveal some type of consistent and causal relationship that F and E have in common with the remaining variables presented in these Tables.

Gunderson and Nelson Data

For exploratory purposes only, the CLPC technique was applied to data collected by Gunderson and Nelson (1962; 1965). This data was collected from six small groups of scientists and Navy personnel living and working together in complete physical isolation from the rest of the world for approximately twelve months at scientific stations on the Antarctic continent.

These men completed an attitude questionnaire both at mid-winter after six months of living in the Antarctic, and at post-winter approximately four months later. The attitude questionnaire was made up of five "attitude study" variables and five "group behavior" variables. The five attitude study variables are:

PA = Physical Adjustment: feelings of adjustment to climatic and living conditions.

EM = Expedition Motivation: expressed interest in remaining or returning on Antarctic expeditions.

TO = Trust in Organization: confidence and trust in general authority and expedition organization.

PU = Personal Usefulness: feeling that personal job is important and that personal gain will be derived from participation.

B = Boredom: expressed boredom, lack of things to do, time drag.

The five group behavior variables are:

C = Compatability: perception of group members and mutually congenial and preferred as personal friends.

T = Teamwork: perception of group members as cooperative and each carrying his share of the work.

GE = Group Efficiency: perception of group as well organized, having concise goals, and scheduled activities.

GA = Group Achievement: perception of group accomplishment and members' pride in same.

DA = Democratic Atmosphere: perception of mutual respect, status leveling, and democratic procedures within group.

Table 6 presents the intercorrelations, including the CLPCs, for all of the above variables measured from responses of 110 men pooled from all six group. The "all-test" $\underline{p}$ level for this matrix, which contains 45 Cross-Lagged comparisons, was computed in the following manner:

$$\frac{1}{45 \times 1 \times 100} \text{ or } p = \frac{.00022222}{2} = .00011111 \text{ (two-tailed test)}$$

$$z' = 3.49 \text{ for } p < .01$$

$$\frac{1}{45 \times 1 \times 20} = \frac{.00111111}{2} = .00055555 \text{ (two-tailed test)}$$

$$z' = 3.26 \text{ for } p < .05$$

As is seen in Table 6, no differences between CLPCs are significant at either the .01 or .05 levels for the conservative, all-test p criteria. Using the standard signle-test p criteria, two significant differences are found between CLPCs. The first is $PA_1GE_2 > PA_2GE_1$ and this difference is significant at $p < .01$ for t_r and $p < .05$ for $t_{z'}$. Both correlations are negative. This relation indicates that the preponderance of causal direction lies from PA to GE and that high prior degree of physical adjustment leads to a low subsequent degree of group efficiency. The second significant difference is $T_1GA_2 > T_2GA_1$ which is significant at the $p < .05$ level for t_r only. Both correlations are relatively large and since they are both positive in sign, the relationship indicates that the higher the degree of prior teamwork, the higher the subsequent degree of group achievement with the preponderance of cause lying in the direction of T to GA.

Summary

The Cross-Lagged Panel Correlation, a quasi-experimental method of analysis presented by Campbell and Stanley (1963) was applied to three sets of data. The method inspects time-lagged relationships between variables. One of the sets of data contained correlation matrices generated from random numbers. The CLPC technique was applied to this data for the purpose of testing some of the difficulties encountered when making tests of significance with multiple comparisons. The methods of correction used seemed to result in an underestimation of significance, i.e., the expected *p* values were uniformly too high.

The next set of data analyzed consisted of attitude measurements collected at Vassar College. Questionnaires were given twice to three separate graduating classes, once in the freshmen year and once during the senior year. Two of the scales presented were the E scale and the F scale. Evidence was presented supporting two conflicting hypotheses concerning the causal relationship between E and F. One hypothesis was that prior authoritarianism lead to subsequent ethnocentrism whereas the alternative hypothesis stated that the greater preponderance of cause was in the other direction, i.e., prior ethnocentrism is a greater cause of subsequent authoritarianism than vice versa. Analysis by means of the CLPC produced no conclusive results, but slightly tended to favor the prior A causes subsequent E hypothesis. Significant differences between CLPCs involving the other variables are also presented.

The third set of data consisted of attitude scores obtained from small groups of men stationed in the Antarctic. The same questionnaire was presented twice, once after the men had been at their stations for six months and again four months later. No interpretable results were obtained because of lack of significant differences between the CLPCs involved.

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., Sanford, R. N. *The Authoritarian Personality*. New York: Harper, 1950.

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.) *Problems in measuring change*. Madison, Wis.: University of Wisconsin Press, 1963, pp. 3-20.

Campbell, D. T. From description to experimentation: interpreting trends as quasi-experiments. In C. W. Harris,(Ed.) *Problems in measuring change*. Madison, Wis.: University of Wisconsin Press, 1963, pp. 212-242.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.) *Handbook of research on teaching*. Chicago: Rand McNally, 1963, pp. 171-246.

Campbell, D. T., Miller, N., Lubetsky, J., & O'Connell, E. J. Varieties of projection and trait attribution. *Psychological Monographs*, 1964, 78, 1-33.

Gunderson, E. K. E., & Nelson, P. D. Attitude changes in small groups under prolonged isolation. Report No. 62-2, U.S. N. Med. N.P. Res. Unit San Diego, March, 1962.

Gunderson, E. K. E., & Nelson, P. D. Measurement of group effectiveness in natural isolated groups. *Journal of Social Psychology*, 1965, 66, 241-249.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. *Experiments on Mass Communication*. Vol. III, Princeton: Princeton University Press, 1949.

Pearson, K., & Filon, L. N. G. Mathematical contributions to the theory of evolution. Trans. Roy. Soc. (London), Series A, Vol. 191. Pp. 259-262. Cited by C. C. Peters & W. R. Van Voorhis in *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.

Rozelle, R. M. An exploration of two quasi-experimental designs: the cross-lagged panel correlation and the multiple time series. Unpublished Master's Thesis, Northwestern University, 1965.

Ryan, T. A. Significance tests for multiple comparison of proportions, variances, and other statistics. *Psychological Bulletin*, 1960, 57, 318-328.

Webster, H., Freedman, M., & Heist, P. The effects of college education. In R. N. Sanford (Ed.) *The American college*. New York: John Wiley, 1961, pp. 811-847.

Table 1. Illustrative Cross-Lagged Panel Correlations

for Montecarlo Data

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_T	Single Test p	All Test p	t_S'	Single Test p	All Test p
A_1A_2	B_1B_2	A_1B_1	A_2B_2	A_1B_2	A_2B_1						
.46	.38	.46	.46	.51	.49	.21			.19		
.58	.56	.46	.62	.40	.33	.35			.47		
.70	.88	.88	.87	.91	.83	2.87	.01		2.36		

Table 2. Cross-Lagged Panel Correlations for Vassar College Class of 1937 (N = 260)

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		Z_r	Single Test P	All Test P	Z_r''	Single Test P	All Test P
F_1F_2 .35	E_1E_2 .62	F_1E_1 .62	F_2E_2 .53	F_1E_2 .45	F_2E_1 .54	1.74			1.40		
SI_1SI_2 .64	DC_1DC_2 .65	SI_1DC_1 .67	SI_2DC_2 .63	SI_1DC_2 .51	SI_2DC_1 .40	2.03	.05		1.57		
SI_1SI_2 .64	MR_1MR_2 .48	SI_1MR_1 .06	SI_2MR_2 .05	SI_1MR_2 .04	SI_2MR_1 .02	.27			.22		
SI_1SI_2 .64	RS_1RS_2 .64	SI_1RS_1 .84	SI_2RS_2 .82	SI_1RS_2 .58	SI_2RS_1 .57	.24			.16		
SI_1SI_2 .64	DS_1DS_2 .65	SI_1DS_1 -.45	SI_2DS_2 -.57	SI_1DS_2 -.31	SI_2DS_1 -.45	2.35	.05		1.86		
SI_1SI_2 .64	SM_1SM_2 .68	SI_1SM_1 .20	SI_2SM_2 .05	SI_1SM_2 .16	SI_2SM_1 -.04	3.07	.01		2.28	.05	
SI_1SI_2 .64	IE_1IE_2 .63	SI_1IE_1 -.57	SI_2IE_2 -.68	SI_1IE_2 -.43	SI_2IE_1 -.47	.72			.56		
SI_1SI_2 .64	F_1F_2 .35	SI_1F_1 -.13	SI_2F_2 .02	SI_1F_2 -.08	SI_2F_1 .05	1.68			1.47		
SI_1SI_2 .64	E_1E_2 .62	SI_1E_1 -.01	SI_2E_2 .17	SI_1E_2 .01	SI_2E_1 .12	1.61			1.25		
DC_1DC_2 .65	MR_1MR_2 .48	DC_1MR_1 .30	DC_2MR_2 .28	DC_1MR_2 .21	DC_2MR_1 .10	1.56			1.27		
DC_1DC_2 .65	RS_1RS_2 .64	DC_1RS_1 .57	DC_2RS_2 .55	DC_1RS_2 .35	DC_2RS_1 .43	1.37			1.07		
DC_1DC_2 .65	DS_1DS_2 .65	DC_1DS_1 -.07	DC_2DS_2 -.18	DC_1DS_2 -.07	DC_2DS_1 -.18	1.67			1.26		

Table 2. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z*	Single Test P	All Test P
DC_1DC_2	SM_1SM_2	DC_1SM_1	DC_2SM_2	DC_1SM_2	DC_2SM_1						
.65	.68	.31	.25	.22	.11	1.72			1.28		
DC_1DC_2	IE_1IE_2	DC_1IE_1	DC_2IE_2	DC_1IE_2	DC_2IE_1						
.65	.63	-.04	-.17	-.09	-.12	.44			.34		
DC_1DC_2	F_1F_2	DC_1F_1	DC_2F_2	DC_1F_2	DC_2F_1						
.65	.35	-.18	-.05	-.11	-.09	.26			.22		
DC_1DC_2	E_1E_2	DC_1E_1	DC_2E_2	DC_1E_2	DC_2E_1						
.65	.62	-.11	.00	.10	-.08	.29			.22		
MR_1MR_2	RS_1RS_2	MR_1RS_1	MR_2RS_2	MR_1RS_2	MR_2RS_1						
.48	.64	.20	.14	.10	.14	.55			.46		
MR_1MR_2	DS_1DS_2	MR_1DS_1	MR_2DS_2	MR_1DS_2	MR_2DS_1						
.48	.65	.27	.22	.12	.15	.42			.34		
MR_1MR_2	SM_1SM_2	MR_1SM_1	MR_2SM_2	MR_1SM_2	MR_2SM_1						
.48	.68	.26	.27	.15	.18	.42			.34		
MR_1MR_2	IE_1IE_2	MR_1IE_1	MR_2IE_2	MR_1IE_2	MR_2IE_1						
.48	.63	.22	.27	.15	.16	.13			.11		
MR_1MR_2	F_1F_2	MR_1F_1	MR_2F_2	MR_1F_2	MR_2F_1						
.48	.35	-.07	-.10	.00	-.07	.87			.79		
MR_1MR_2	E_1E_2	MR_1E_1	MR_2E_2	MR_1E_2	MR_2E_1						
.48	.62	-.15	-.05	-.07	-.12	.68			.57		
RS_1RS_2	DS_1DS_2	RS_1DS_1	RS_2DS_2	RS_1DS_2	RS_2DS_1						
.64	.65	-.32	-.43	-.24	-.33	1.42			1.11		
RS_1RS_2	SM_1SM_2	RS_1SM_1	RS_2SM_2	RS_1SM_2	RS_2SM_1						
.64	.68	.34	.27	.26	.10	2.50	.05		1.87		
RS_1RS_2	IE_1IE_2	RS_1IE_1	RS_2IE_2	RS_1IE_2	RS_2IE_1						
.64	.63	-.57	-.70	-.44	-.45	.18			.14		

Table 2 Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
RS_1RS_2	F_1F_2	RS_1F_1	RS_2F_2	RS_1F_2	RS_2F_1						
.64	.35	-.23	-.06	-.13	-.05	1.04			.91		
RS_1RS_2	E_1E_2	RS_1E_1	RS_2E_2	RS_1E_2	RS_2E_1						
.64	.62	-.10	.00	-.07	-.02	.73			.56		
DS_1DS_2	SM_1SM_2	DS_1SM_1	DS_2SM_2	DS_1SM_2	DS_2SM_1						
.65	.68	.63	.61	.38	.50	2.19	.05		1.69		
DS_1DS_2	IE_1IE_2	DS_1IE_1	DS_2IE_2	DS_1IE_2	DS_2IE_1						
.65	.63	.72	.77	.55	.51	.84			.63		
DS_1DS_2	F_1F_2	DS_1F_1	DS_2F_2	DS_1F_2	DS_2F_1						
.65	.35	-.45	-.23	-.11	-.44	4.62	.01	.01	4.10	.01	.01
DS_1DS_2	E_1E_2	DS_1E_1	DS_2E_2	DS_1E_2	DS_2E_1						
.65	.62	-.44	-.45	-.31	-.43	1.97	.05		1.57		
SM_1SM_2	IE_1IE_2	SM_1IE_1	SM_2IE_2	SM_1IE_2	SM_2IE_1						
.68	.63	.24	.27	.24	.15	1.39			1.06		
SM_1SM_2	F_1F_2	SM_1F_1	SM_2F_2	SM_1F_2	SM_2F_1						
.68	.35	-.72	-.38	-.30	-.57	4.22	.01	.01	3.84	.01	.05
SM_1SM_2	E_1E_2	SM_1E_1	SM_2E_2	SM_1E_2	SM_2E_1						
.68	.62	-.53	-.55	-.41	-.52	1.98	.05				
IE_1IE_2	F_1F_2	IE_1F_1	IE_2F_2	IE_1F_2	IE_2F_1						
.63	.35	-.13	-.10	-.02	-.15	1.69			1.48		
IE_1IE_2	E_1E_2	IE_1E_1	IE_2E_2	IE_1E_2	IE_2E_1						
.63	.62	-.23	-.23	-.16	-.23	1.05			.82		

Table 3. Cross-Lagged Panel Correlations for Vassar College Class of 1958 (N = 249)

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test *P*	All Test *P*	t_z'	Single Test *P*	All Test *P*
F_1F_2	E_1E_2	F_1E_1	F_2E_2	F_1E_2	F_2E_1						
.54	.65	.60	.75	.48	.48	.00			.00		
SI_1SI_2	DC_1DC_2	SI_1DC_1	SI_2DC_2	SI_1DC_2	SI_2DC_1						
.63	.57	.59	.61	.39	.39	.00			.00		
SI_1SI_2	MR_1MR_2	SI_1MR_1	SI_2MR_2	SI_1MR_2	SI_2MR_1						
.63	.49	.09	.03	.03	-.03	.80			.66		
SI_1SI_2	RS_1RS_2	SI_1RS_1	SI_2RS_2	SI_1RS_2	SI_2RS_1						
.63	.63	.84	.85	.55	.56	.24			.16		
SI_1SI_2	DS_1DS_2	SI_1DS_1	SI_2DS_2	SI_1DS_2	SI_2DS_1						
.63	.66	-.44	-.49	-.30	-.39	1.44			1.13		
SI_1SI_2	SM_1SM_2	SI_1SM_1	SI_2SM_2	SI_1SM_2	SI_2SM_1						
.63	.62	.23	.16	.15	.02	1.88			1.45		
SI_1SI_2	IE_1IE_2	SI_1IE_1	SI_2IE_2	SI_1IE_2	SI_2IE_1						
.63	.66	-.63	-.65	-.39	-.51	2.15	.05		1.67		
SI_1SI_2	F_1F_2	SI_1F_1	SI_2F_2	SI_1F_2	SI_2F_1						
.63	.54	.01	-.04	-.04	.05	1.23			.99		
SI_1SI_2	E_1E_2	SI_1E_1	SI_2E_2	SI_1E_2	SI_2E_1						
.63	.65	-.10	.03	-.06	.04	1.45			1.11		
DC_1DC_2	MR_1MR_2	DC_1MR_1	DC_2MR_2	DC_1MR_2	DC_2MR_1						
.57	.49	.38	.41	.22	.21	.14			.11		
DC_1DC_2	RS_1RS_2	DC_1RS_1	DC_2RS_2	DC_1RS_2	DC_2RS_1						
.57	.63	.54	.50	.32	.33	.15			.12		
DC_1DC_2	DS_1DS_2	DC_1DS_1	DC_2DS_2	DC_1DS_2	DC_2DS_1						
.57	.66	-.21	-.05	-.10	-.12	.28			.22		

Table 3. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	$t_{z'}$	Single Test P	All Test P
DC_1DC_2	SM_1SM_2	DC_1SM_1	DC_2SM_2	DC_1SM_2	DC_2SM_1						
.57	.62	.25	.30	.12	.15	.42			.33		
DC_1DC_2	IE_1IE_2	DC_1IE_1	DC_2IE_2	DC_1IE_2	DC_2IE_1						
.57	.66	-.09	-.04	-.07	-.06	.14			.11		
DC_1DC_2	F_1F_2	DC_1F_1	DC_2F_2	DC_1F_2	DC_2F_1						
.57	.54	.04	-.09	.05	.05	.00			.00		
DC_1DC_2	E_1E_2	DC_1E_1	DC_2E_2	DC_1E_2	DC_2E_1						
.57	.65	-.04	.03	.06	.03	.42			.33		
MR_1MR_2	RS_1RS_2	MR_1RS_1	MR_2RS_2	MR_1RS_2	MR_2RS_1						
.49	.63	.15	.13	.02	.09	.94			.77		
MR_1MR_2	DS_1DS_2	MR_1DS_1	MR_2DS_2	MR_1DS_2	MR_2DS_1						
.49	.66	.20	.30	.22	.12	1.39			1.14		
MR_1MR_2	SM_1SM_2	MR_1SM_1	MR_2SM_2	MR_1SM_2	MR_2SM_1						
.49	.62	.28	.29	.11	.19	1.10			.90		
MR_1MR_2	IE_1IE_2	MR_1IE_1	MR_2IE_2	MR_1IE_2	MR_2IE_1						
.49	.66	.24	.29	.17	.18	.13			.11		
MR_1MR_2	F_1F_2	MR_1F_1	MR_2F_2	MR_1F_2	MR_2F_1						
.49	.54	-.08	-.09	.00	-.11	1.44			1.22		
MR_1MR_2	E_1E_2	MR_1E_1	MR_2E_2	MR_1E_2	MR_2E_1						
.49	.65	-.08	-.02	.03	-.03	.81			.66		
RS_1RS_2	DS_1DS_2	RS_1DS_1	RS_2DS_2	RS_1DS_2	RS_2DS_1						
.63	.66	-.31	-.34	-.21	-.30	1.37			1.06		
RS_1RS_2	SM_1SM_2	RS_1SM_1	RS_2SM_2	RS_1SM_2	RS_2SM_1						
.63	.62	.40	.34	.24	.15	1.34			1.03		

Table 3. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	$t_{z'}$	Single Test P	All Test P
RS_1RS_2	IE_1IE_2	RS_1IE_1	RS_2IE_2	RS_1IE_2	RS_2IE_1						
.63	.66	-.66	-.67	-.43	-.54	2.05	.05		1.59		
RS_1RS_2	F_1F_2	RS_1F_1	RS_2F_2	RS_1F_2	RS_2F_1						
.63	.54	-.14	-.18	-.10	-.10	.00			.00		
RS_1RS_2	E_1E_2	RS_1E_1	RS_2E_2	RS_1E_2	RS_2E_1						
.63	.65	-.17	-.11	-.10	-.08	.29			.22		
DS_1DS_2	SM_1SM_2	DS_1SM_1	DS_2SM_2	DS_1SM_2	DS_2SM_1						
.66	.62	.40	.63	.37	.47	1.66			1.34		
DS_1DS_2	IE_1IE_2	DS_1IE_1	DS_2IE_2	DS_1IE_2	DS_2IE_1						
.66	.66	.64	.69	.47	.47	.00			.00		
DS_1DS_2	F_1F_2	DS_1F_1	DS_2F_2	DS_1F_2	DS_2F_1						
.66	.54	-.50	-.50	-.34	-.43	1.43			1.17		
DS_1DS_2	E_1E_2	DS_1E_1	DS_2E_2	DS_1E_2	DS_2E_1						
.66	.65	-.35	-.43	-.30	-.35	.79			.62		
SM_1SM_2	IE_1IE_2	SM_1IE_1	SM_2IE_2	SM_1IE_2	SM_2IE_1						
.62	.66	.15	.23	.19	.07	1.76			1.35		
SM_1SM_2	F_1F_2	SM_1F_1	SM_2F_2	SM_1F_2	SM_2F_1						
.62	.54	-.67	-.67	-.47	-.48	.18			.14		
SM_1SM_2	E_1E_2	SM_1E_1	SM_2E_2	SM_1E_2	SM_2E_1						
.62	.65	-.53	-.55	-.43	-.39	.67			.53		
IE_1IE_2	F_1F_2	IE_1F_1	IE_2F_2	IE_1F_2	IE_2F_1						
.66	.54	-.14	-.12	-.01	-.09	1.12			.88		
IE_1IE_2	E_1E_2	IE_1E_1	IE_2E_2	IE_1E_2	IE_2E_1						
.66	.65	-.06	-.11	.00	-.08	1.18			.88		

Table 4. Cross-Lagged Panel Correlations for Vassar College Class of 1959 (N = 247)

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test *P*	All Test *P*	t_z	Single Test *P*	All Test *P*
F_1F_2 .45	E_1E_2 .58	F_1E_1 .61	F_2E_2 .69	F_1E_2 .40	F_2E_1 .41	.16			.13		
SI_1SI_2 .62	DC_1DC_2 .71	SI_1DC_1 .66	SI_2DC_2 .71	SI_1DC_2 .52	SI_2DC_1 .36	2.93	.01		2.20	.05	
SI_1SI_2 .62	MR_1MR_2 .53	SI_1MR_1 .07	SI_2MR_2 .09	SI_1MR_2 .04	SI_2MR_1 -.08	1.64			1.32		
SI_1SI_2 .62	RS_1RS_2 .66	SI_1RS_1 .85	SI_2RS_2 .83	SI_1RS_2 .59	SI_2RS_1 .56	.74			.49		
SI_1SI_2 .62	DS_1DS_2 .66	SI_1DS_1 -.46	SI_2DS_2 -.44	SI_1DS_2 -.28	SI_2DS_1 -.40	1.90			1.50		
SI_1SI_2 .62	SM_1SM_2 .65	SI_1SM_1 .17	SI_2SM_2 .19	SI_1SM_2 .16	SI_2SM_1 .03	1.84			1.45		
SI_1SI_2 .62	IE_1IE_2 .66	SI_1IE_1 -.64	SI_2IE_2 -.63	SI_1IE_2 -.41	SI_2IE_1 -.51	1.79			1.40		
SI_1SI_2 .62	F_1F_2 .45	SI_1F_1 .04	SI_2F_2 .00	SI_1F_2 .04	SI_2F_1 .04	.00			.00		
SI_1SI_2 .62	E_1E_2 .58	SI_1E_1 .01	SI_2E_2 .01	SI_1E_2 -.04	SI_2E_1 .03	.97			.77		
DC_1DC_2 .71	MR_1MR_2 .53	DC_1MR_1 .35	DC_2MR_2 .37	DC_1MR_2 .25	DC_2MR_1 .17	1.17			.92		
DC_1DC_2 .71	RS_1RS_2 .66	DC_1RS_1 .56	DC_2RS_2 .53	DC_1RS_2 .30	DC_2RS_1 .44	2.38	.05		1.79		
DC_1DC_2 .71	DS_1DS_2 .66	DC_1DS_1 -.12	DC_2DS_2 -.10	DC_1DS_2 -.07	DC_2DS_1 -.13	.92			.66		

Table 4. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
DC_1DC_2	SM_1SM_2	DC_1SM_1	DC_2SM_2	DC_1SM_2	DC_2SM_1						
.71	.65	.23	.25	.13	.14	.26			.19		
DC_1DC_2	IE_1IE_2	DC_1IE_1	DC_2IE_2	DC_1IE_2	DC_2IE_1						
.71	.66	-.11	-.09	-.02	-.15	2.00	.05		1.44		
DC_1DC_2	F_1F_2	DC_1F_1	DC_2F_2	DC_1F_2	DC_2F_1						
.71	.45	.05	-.04	.05	.02	.40			.33		
DC_1DC_2	E_1E_2	DC_1E_1	DC_2E_2	DC_1E_2	DC_2E_1						
.71	.58	-.02	-.03	-.06	-.02	.58			.44		
MR_1MR_2	RS_1RS_2	MR_1RS_1	MR_2RS_2	MR_1RS_2	MR_2RS_1						
.53	.66	.19	.17	-.01	.13	1.96			1.55		
MR_1MR_2	DS_1DS_2	MR_1DS_1	MR_2DS_2	MR_1DS_2	MR_2DS_1						
.53	.66	.25	.21	.16	.19	.42			.34		
MR_1MR_2	SM_1SM_2	MR_1SM_1	MR_2SM_2	MR_1SM_2	MR_2SM_1						
.53	.65	.24	.24	.05	.26	2.98	.01		2.38	.05	
MR_1MR_2	IE_1IE_2	MR_1IE_1	MR_2IE_2	MR_1IE_2	MR_2IE_1						
.53	.66	.28	.24	.22	.20	.28			.23		
MR_1MR_2	F_1F_2	MR_1F_1	MR_2F_2	MR_1F_2	MR_2F_1						
.53	.45	.02	-.07	.01	-.01	.25			.22		
MR_1MR_2	E_1E_2	MR_1E_1	MR_2E_2	MR_1E_2	MR_2E_1						
.53	.58	-.05	-.05	-.02	-.07	.66			.55		
RS_1RS_2	DS_1DS_2	RS_1DS_1	RS_2DS_2	RS_1DS_2	RS_2DS_1						
.66	.66	-.29	-.27	-.16	-.28	1.52			1.15		
RS_1RS_2	SM_1SM_2	RS_1SM_1	RS_2SM_2	RS_1SM_2	RS_2SM_1						
.66	.65	.37	.39	.28	.21	1.08			.82		
RS_1RS_2	IE_1IE_2	RS_1IE_1	RS_2IE_2	RS_1IE_2	RS_2IE_1						
.66	.66	-.64	-.64	-.40	-.51	1.98			1.53		

Table 4. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
RS_1RS_2	F_1F_2	RS_1F_1	RS_2F_2	RS_1F_2	RS_2F_1						
.66	.45	-.05	-.19	-.11	-.04	.93			.77		
RS_1RS_2	E_1E_2	RS_1E_1	RS_2E_2	RS_1E_2	RS_2E_1						
.66	.58	-.12	-.15	-.16	-.10	.86			.67		
DS_1DS_2	SM_1SM_2	DS_1SM_1	DS_2SM_2	DS_1SM_2	DS_2SM_1						
.66	.65	.59	.63	.34	.52	3.15	.01		2.45	.05	
DS_1DS_2	IE_1IE_2	DS_1IE_1	DS_2IE_2	DS_1IE_2	DS_2IE_1						
.66	.66	.69	.72	.52	.48	.77			.58		
DS_1DS_2	F_1F_2	DS_1F_1	DS_2F_2	DS_1F_2	DS_2F_1						
.66	.45	-.39	-.54	-.40	-.37	.45			.38		
DS_1DS_2	E_1E_2	DS_1E_1	DS_2E_2	DS_1E_2	DS_2E_1						
.66	.58	-.34	-.42	-.30	-.31						
SM_1SM_2	IE_1IE_2	SM_1IE_1	SM_2IE_2	SM_1IE_2	SM_2IE_1						
.65	.66	.18	.21	.21	.03	2.70	.01		2.02	.05	
SM_1SM_2	F_1F_2	SM_1F_1	SM_2F_2	SM_1F_2	SM_2F_1						
.65	.45	-.45	-.68	-.50	-.43	1.15			.98		
SM_1SM_2	E_1E_2	SM_1E_1	SM_2E_2	SM_1E_2	SM_2E_1						
.65	.58	-.44	-.55	-.48	-.38	1.65			1.35		
IE_1IE_2	F_1F_2	IE_1F_1	IE_2F_2	IE_1F_2	IE_2F_1						
.66	.45	-.07	-.15	-.08	-.11	.40			.33		
IE_1IE_2	E_1E_2	IE_1E_1	IE_2E_2	IE_1E_2	IE_2E_1						
.66	.58	-.06	-.12	-.06	-.08	.28			.22		

Table 5. Cross-Lagged Panel Correlations for Vassar College All Classes Combined (N = 756)

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	$t_{z'}$	Single Test P	All Test P
F_1F_2	E_1E_2	F_1E_1	F_2E_2	F_1E_2	F_2E_1						
.41	.61	.63	.62	.44	.38	1.71			1.40		
SI_1SI_2	DC_1DC_2	SI_1DC_1	SI_2DC_2	SI_1DC_2	SI_2DC_1						
.63	.64	.63	.63	.47	.39	2.44	.05		1.90		
SI_1SI_2	MR_1MR_2	SI_1MR_1	SI_2MR_2	SI_1MR_2	SI_2MR_1						
.63	.50	.07	.06	.03	-.04	1.65			1.35		
SI_1SI_2	RS_1RS_2	SI_1RS_1	SI_2RS_2	SI_1RS_2	SI_2RS_1						
.63	.64	.84	.83	.57	.56	.42			.28		
SI_1SI_2	DS_1DS_2	SI_1DS_1	SI_2DS_2	SI_1DS_2	SI_2DS_1						
.63	.66	-.45	-.50	-.30	-.42	3.39	.01		2.68	.01	
SI_1SI_2	SM_1SM_2	SI_1SM_1	SI_2SM_2	SI_1SM_2	SI_2SM_1						
.63	.65	.19	.13	.16	-.01	4.37	.01		3.32	.01	
SI_1SI_2	IE_1IE_2	SI_1IE_1	SI_2IE_2	SI_1IE_2	SI_2IE_1						
.63	.65	-.61	-.66	-.42	-.50	2.50	.05		1.97		
SI_1SI_2	F_1F_2	SI_1F_1	SI_2F_2	SI_1F_2	SI_2F_1						
.63	.41	-.01	.00	-.04	.06	2.26	.05		1.94		
SI_1SI_2	E_1E_2	SI_1E_1	SI_2E_2	SI_1E_2	SI_2E_1						
.63	.61	-.02	.07	-.03	.08	2.73	.01		2.13	.05	
DC_1DC_2	MR_1MR_2	DC_1MR_1	DC_2MR_2	DC_1MR_2	DC_2MR_1						
.64	.50	.34	.35	.23	.16	1.72			1.41		
DC_1DC_2	RS_1RS_2	DC_1RS_1	DC_2RS_2	DC_1RS_2	DC_2RS_1						
.64	.64	.55	.52	.32	.40	2.28	.05		1.78		
DC_1DC_2	DS_1DS_2	DC_1DS_1	DC_2DS_2	DC_1DS_2	DC_2DS_1						
.64	.66	-.14	-.11	-.08	-.14	1.54			1.17		

Table 5. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	$t_{z'}$	Single Test P	All Test P
DC_1DC_2	SM_1SM_2	DC_1SM_1	DC_2SM_2	DC_1SM_2	DC_2SM_1						
.64	.65	.26	.27	.15	.13	.51			.39		
DC_1DC_2	IE_1IE_2	DC_1IE_1	DC_2IE_2	DC_1IE_2	DC_2IE_1						
.64	.65	-.08	-.10	-.06	-.11	1.27			.97		
DC_1DC_2	F_1F_2	DC_1F_1	DC_2F_2	DC_1F_2	DC_2F_1						
.64	.41	-.03	-.06	-.02	-.01	.22			.19		
DC_1DC_2	E_1E_2	DC_1E_1	DC_2E_2	DC_1E_2	DC_2E_1						
.64	.61	-.03	-.02	-.03	-.02	.24			.19		
MR_1MR_2	RS_1RS_2	MR_1RS_1	MR_2RS_2	MR_1RS_2	MR_2RS_1						
.50	.64	.17	.15	.03	.12	2.14	.05		1.75		
MR_1MR_2	DS_1DS_2	MR_1DS_1	MR_2DS_2	MR_1DS_2	MR_2DS_1						
.50	.66	.25	.24	.17	.15	.48			.39		
MR_1MR_2	SM_1SM_2	MR_1SM_1	MR_2SM_2	MR_1SM_2	MR_2SM_1						
.50	.65	.26	.27	.10	.20	2.43	.05		1.98	.05	
MR_1MR_2	IE_1IE_2	MR_1IE_1	MR_2IE_2	MR_1IE_2	MR_2IE_1						
.50	.65	.25	.26	.19	.18	.24			.20		
MR_1MR_2	F_1F_2	MR_1F_1	MR_2F_2	MR_1F_2	MR_2F_1						
.50	.41	-.06	-.08	.00	-.06	1.31			1.16		
MR_1MR_2	E_1E_2	MR_1E_1	MR_2E_2	MR_1E_2	MR_2E_1						
.50	.61	-.11	-.04	-.02	-.08	1.40			1.16		
RS_1RS_2	DS_1DS_2	RS_1DS_1	RS_2DS_2	RS_1DS_2	RS_2DS_1						
.64	.66	-.31	-.35	-.21	-.30	2.41	.05		1.86		
RS_1RS_2	SM_1SM_2	RS_1SM_1	RS_2SM_2	RS_1SM_2	RS_2SM_1						
.64	.65	.36	.33	.26	.15	2.91	.01		2.23	.05	
RS_1RS_2	IE_1IE_2	RS_1IE_1	RS_2IE_2	RS_1IE_2	RS_2IE_1						
.64	.65	-.62	-.67	-.43	-.50	2.21	.05		1.73		

Table 5. Continued

Test-Retest		Synchronous Cross Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
RS_1RS_2	F_1F_2	RS_1F_1	RS_2F_2	RS_1F_2	RS_2F_1						
.64	.41	-.13	-.12	-.12	-.06	1.36			1.17		
RS_1RS_2	E_1E_2	RS_1E_1	RS_2E_2	RS_1E_2	RS_2E_1						
.64	.61	-.12	-.08	-.11	-.05	1.50			1.17		
DS_1DS_2	SM_1SM_2	DS_1SM_1	DS_2SM_2	DS_1SM_2	DS_2SM_1						
.66	.65	.60	.63	.37	.50	4.00	.01	.01	3.12	.01	
DS_1DS_2	IE_1IE_2	DS_1IE_1	DS_2IE_2	DS_1IE_2	DS_2IE_1						
.66	.65	.68	.73	.52	.49	1.02			.78		
DS_1DS_2	F_1F_2	DS_1F_1	DS_2F_2	DS_1F_2	DS_2F_1						
.66	.41	-.46	-.38	-.25	-.42	4.31	.01	.01	3.73	.01	.05
DS_1DS_2	E_1E_2	DS_1E_1	DS_2E_2	DS_1E_2	DS_2E_1						
.66	.61	-.40	-.44	-.30	-.37	1.92			1.53		
SM_1SM_2	IE_1IE_2	SM_1IE_1	SM_2IE_2	SM_1IE_2	SM_2IE_1						
.65	.65	.20	.24	.22	.09	3.39	.01	.05	2.58	.05	
SM_1SM_2	F_1F_2	SM_1F_1	SM_2F_2	SM_1F_2	SM_2F_1						
.65	.41	-.62	-.53	-.39	-.49	2.83	.01		2.41	.05	
SM_1SM_2	E_1E_2	SM_1E_1	SM_2E_2	SM_1E_2	SM_2E_1						
.65	.61	-.51	-.55	-.44	-.43	.29			.23		
IE_1IE_2	F_1F_2	IE_1F_1	IE_2F_2	IE_1F_2	IE_2F_1						
.65	.41	-.12	-.12	-.03	-.13	2.29	.05		1.95		
IE_1IE_2	E_1E_2	IE_1E_1	IE_2E_2	IE_1E_2	IE_2E_1						
.65	.61	-.13	-.15	-.08	-.15	1.77			1.37		

Table 6. Cross-Lagged Panel Correlations for the Gunderson and Nelson Antarctic Study (N = 110)

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
PA_1PA_2	EM_1EM_2	PA_1EM_1	PA_2EM_2	PA_1EM_2	PA_2EM_1						
.63	.78	.27	.14	.19	.18	.11			.07		
PA_1PA_2	TO_1TO_2	PA_1TO_1	PA_2TO_2	PA_1TO_2	PA_2TO_1						
.63	.68	.10	.05	-.04	.10	1.38			1.02		
PA_1PA_2	PU_1PU_2	PA_1PU_1	PA_2PU_2	PA_1PU_2	PA_2PU_1						
.63	.69	.21	.09	.09	.10	.09			.07		
PA_1PA_2	B_1B_2	PA_1B_1	PA_2B_2	PA_1B_2	PA_2B_1						
.63	.58	-.08	-.14	-.11	-.12	.09			.07		
PA_1PA_2	C_1C_2	PA_1C_1	PA_2C_2	PA_1C_2	PA_2C_1						
.63	.61	.06	.17	.03	.02	.09			.07		
PA_1PA_2	T_1T_2	PA_1T_1	PA_2T_2	PA_1T_2	PA_2T_1						
.63	.61	-.12	.14	-.05	-.06	.09			.07		
PA_1PA_2	GE_1GE_2	PA_1GE_1	PA_2GE_2	PA_1GE_2	PA_2GE_1						
.63	.72	-.14	-.02	-.32	-.04	2.99	.01		2.13	.05	
PA_1PA_2	GA_1GA_2	PA_1GA_1	PA_2GA_2	PA_1GA_2	PA_2GA_1						
.63	.65	-.11	.01	-.10	-.13	.29			.22		
PA_1PA_2	DA_1DA_2	PA_1DA_1	PA_2DA_2	PA_1DA_2	PA_2DA_1						
.63	.67	-.08	.05	.02	-.08	.97			.73		
EM_1EM_2	TO_1TO_2	EM_1TO_1	EM_2TO_2	EM_1TO_2	EM_2TO_1						
.78	.68	.26	.47	.33	.29	.45			.32		
EM_1EM_2	PU_1PU_2	EM_1PU_1	EM_2PU_2	EM_1PU_2	EM_2PU_1						
.78	.69	.65	.62	.51	.53	.26			.20		
EM_1EM_2	B_1B_2	EM_1B_1	EM_2B_2	EM_1B_2	EM_2B_1						
.78	.58	-.40	-.33	-.39	-.35	.44			.33		

Table 6. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
EM_1EM_2 .78	C_1C_2 .61	EM_1C_1 .38	EM_2C_2 .51	EM_1C_2 .39	EM_2C_1 .33	.66			.50		
EM_1EM_2 .78	T_1T_2 .61	EM_1T_1 .29	EM_2T_2 .38	EM_1T_2 .36	EM_2T_1 .24	1.30			.96		
EM_1EM_2 .78	GE_1GE_2 .72	EM_1GE_1 .17	EM_2GE_2 .12	EM_1GE_2 .07	EM_2GE_1 .10	.33			.22		
EM_1EM_2 .78	GA_1GA_2 .65	EM_1GA_1 .24	EM_2GA_2 .23	EM_1GA_2 .22	EM_2GA_1 .13	.96			.67		
EM_1EM_2 .78	DA_1DA_2 .67	EM_1DA_1 .20	EM_2DA_2 .45	EM_1DA_2 .31	EM_2DA_1 .22	.98			.70		
TO_1TO_2 .68	PU_1PU_2 .69	TO_1PU_1 .50	TO_2PU_2 .61	TO_1PU_2 .37	TO_2PU_1 .46	1.04			.79		
TO_1TO_2 .68	B_1B_2 .58	TO_1B_1 -.20	TO_2B_2 -.16	TO_1B_2 -.10	TO_2B_1 -.27	1.67			1.31		
TO_1TO_2 .68	C_1C_2 .61	TO_1C_1 .49	TO_2C_2 .40	TO_1C_2 .28	TO_2C_1 .42	1.49			1.17		
TO_1TO_2 .68	T_1T_2 .61	TO_1T_1 .40	TO_2T_2 .42	TO_1T_2 .26	TO_2T_1 .44	1.92			1.50		
TO_1TO_2 .68	GE_1GE_2 .72	TO_1GE_1 .25	TO_2GE_2 .23	TO_1GE_2 .15	TO_2GE_1 .24	.95			.68		
TO_1TO_2 .68	GA_1GA_2 .65	TO_1GA_1 .41	TO_2GA_2 .39	TO_1GA_2 .27	TO_2GA_1 .34	.74			.56		
TO_1TO_2 .68	DA_1DA_2 .67	TO_1DA_1 .46	TO_2DA_2 .57	TO_1DA_2 .43	TO_2DA_1 .45	.23			.18		
PU_1PU_2 .69	B_1B_2 .58	PU_1B_1 -.52	PU_2B_2 -.34	PU_1B_2 -.39	PU_2B_1 -.41	.21			.17		

Table 6. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z'	Single Test P	All Test P
PU_1PU_2	G_1G_2	PU_1C_1	PU_2C_2	PU_1C_2	PU_2C_1						
.69	.61	.47	.47	.39	.34	.54			.42		
PU_1PU_2	T_1T_2	PU_1T_1	PU_2T_2	PU_1T_2	PU_2T_1						
.69	.61	.36	.39	.37	.32	.53			.41		
PU_1PU_2	GE_1GE_2	PU_1GE_1	PU_2GE_2	PU_1GE_2	PU_2GE_1						
.69	.72	.21	.16	.11	.19	.84			.59		
PU_1PU_2	GA_1GA_2	PU_1GA_1	PU_2GA_2	PU_1GA_2	PU_2GA_1						
.69	.65	.39	.32	.32	.26	.63			.47		
PU_1PU_2	DA_1DA_2	PU_1DA_1	PU_2DA_2	PU_1DA_2	PU_2DA_1						
.69	.67	.29	.53	.38	.32	.65			.50		
B_1B_2	C_1C_2	B_1C_1	B_2C_2	B_1C_2	B_2C_1						
.58	.61	-.33	-.33	-.31	-.33	.20			.16		
B_1B_2	T_1T_2	B_1T_1	B_2T_2	B_1T_2	B_2T_1						
.58	.61	-.24	-.31	-.32	-.25	.68			.55		
B_1B_2	GE_1GE_2	B_1GE_1	B_2GE_2	B_1GE_2	B_2GE_1						
.58	.72	-.13	-.01	-.08	-.10	.19			.14		
B_1B_2	GA_1GA_2	B_1GA_1	B_2GA_2	B_1GA_2	B_2GA_1						
.58	.65	-.16	-.14	-.21	-.14	.67			.52		
B_1B_2	DA_1DA_2	B_1DA_1	B_2DA_2	B_1DA_2	B_2DA_1						
.58	.67	-.14	-.17	-.25	-.22						
C_1C_2	T_1T_2	C_1T_1	C_2T_2	C_1T_2	C_2T_1						
.61	.61	.76	.80	.53	.52	.14			.10		
C_1C_2	GE_1GE_2	C_1GE_1	C_2GE_2	C_1GE_2	C_2GE_1						
.61	.72	.54	.53	.50	.39	1.28			1.00		
C_1C_2	GA_1GA_2	C_1GA_1	C_2GA_2	C_1GA_2	C_2GA_1						
.61	.65	.71	.69	.50	.44	.74			.56		

Table 6. Continued

Test-Retest		Synchronous Cross-Variable		Cross-Lagged Correlations		t_r	Single Test P	All Test P	t_z	Single Test P	All Test P
C_1C_2 .61	DA_1DA_2 .67	C_1DA_1 .56	C_2DA_2 .51	C_1DA_2 .40	C_2DA_1 .37	.32			.25		
T_1T_2 .61	GE_1GE_2 .72	T_1GE_1 .57	T_2GE_2 .65	T_1GE_2 .55	T_2GE_1 .44	1.35			1.06		
T_1T_2 .61	GA_1GA_2 .65	T_1GA_1 .75	T_2GA_2 .78	T_1GA_2 .63	T_2GA_1 .49	2.01	.05		1.50		
T_1T_2 .61	DA_1DA_2 .67	T_1DA_1 .52	T_2DA_2 .45	T_1DA_2 .36	T_2DA_1 .36	.00			.00		
GE_1GE_2 .72	GA_1GA_2 .65	GE_1GA_1 .63	GE_2GA_2 .66	GE_1GA_2 .56	GE_2GA_1 .57	.13			.10		
GE_1GE_2 .72	DA_1DA_2 .67	GE_1DA_1 .37	GE_2DA_2 .24	GE_1DA_2 .22	GE_2DA_1 .31	.96			.70		
GA_1GA_2 .65	DA_1DA_2 .67	GA_1DA_1 .42	GA_2DA_2 .36	GA_1DA_2 .24	GA_2DA_1 .31	.73			.55		

A STUDY OF THE EFFECT OF PROXIMALLY AUTOCORRELATED ERROR ON TESTS OF SIGNIFICANCE FOR THE INTERRUPTED TIME SERIES QUASI-EXPERIMENTAL DESIGN*

Joyce Sween and Donald T. Campbell
August 1965
Northwestern university

---

*Supported by Project C998, Contract 3-20-001, Educational Media Branch, Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

A Study of the Effect of Proximally Autocorrelated Error on Tests of Significance for the Interrupted Time-Series Quasi-Experimental Design

Joyce Sween and Donald T. Campbell
Northwestern University

The time series experiment has long been a common research design in the biological and physical sciences. However, with psychological and sociological data certain problems of analysis and interpretation occur when the interrupted time series design is employed (Campbell, 1963; Campbell & Stanley, 1962; Holtzman, 1963).

One of the problems which is of particular concern to the social scientist (for whom the magnitude and clarity of effects is not always as clear-cut as in the biological and physical sciences) is testing for the significance of the change, X. It is desired to have some statistical test of significance that would distinguish the effect of an intervening event or experimental variable from purely random fluctuation. Although there are no tests of significance that are completely suitable to the time series situation, several possibilities for such a test do exist and merit consideration (Campbell, 1963; Campbell & Stanley, 1962).

Another problem of fundamental concern to the social scientist is the possibility of sequential dependency in successive observations of the time series. Measurements made at time points which are closer together may be more strongly related than measurements made farther apart in time. Statistical models on which the tests of significance are based generally assume that the observed values exhibit independent error. Thus, the significant autocorrelation which exists when a given observation is dependent upon the observations preceding it violates the assumption of independence in these tests. Although this circumstance does not prevent estimation (by least square methods) of the regression coefficients which the various tests of significance use, the classical formulas for the mean errors of the estimated coefficients are in actuality no longer valid. When successive observations through time are correlated, perhaps, to a lag of several points, use of existing tests of significance may be inappropriate.

The primary purpose of the present study was to investigate the appropriateness of several tests of significance for use with interrupted time series data. The three main tests were those described by Mood (1950, pp. 297-298; pp. 350-358) and Walker and Lev (1953, pp. 390-395; pp. 399-400).

The interrupted time series experiment as considered in the present study consists of a finite sequence of observations which are real-valued measurements of some individual or group obtained at n successive equally spaced points in time. An experimental change is then introduced or a major change of conditions occurs at some point in the time series. In the present study, the experimental change or treatment is assumed to occur midway between two consecutive measurements. It can be diagrammed as follows:

Interrupted Time Series Design

$$\begin{array}{lccccccccc} \text{Observation at time } t_i: & y_1 & y_2 & \cdots & y_m & & y_{m+1} & y_{m+2} & \cdots & y_n \\ \text{Time:} & t_1 & t_2 & \cdots & t_m & \uparrow & t_{m+1} & t_{m+2} & \cdots & t_n \\ & & & & & \text{experimental change} & & & & \end{array}$$

The y_i are the measurements recorded as time t_i. The values $y_i (i = 1,2, \ldots m)$ obtained prior to the treatment are referred to as pre-change values. The values y_i ($i = m + 1, m + 2, \ldots n$) obtained after the treatment are post-change values. If the treatment has produced an effect, a discontinuity of measurements is recorded in the time series. The statistical tests of significance which were investigated in the present study as possible techniques for distinguishing such a treatment effect from purely random fluctuation are described in detail below:

(a) Mood Test. This is a *t* test (Mood, 1950, pp. 297-298) for the significance of the first post-change observation from a value predicted by a linear fit of the pre-change observations. The formula for *t* is

$$t = \frac{(y_0 - \hat{y}_0)/\sigma_u}{\sqrt{m\hat{\sigma}^2/(m-2)\sigma^2}}$$

The estimate, $\hat{y}_0$, of the first post-change value is obtained from the pre-change regression estimates

$$\hat{\beta} = \frac{\sum_{i=1}^{m} t_i y_i - \left(\sum_{i=1}^{m} t_i\right)\left(\sum_{i=1}^{m} y_i\right)/m}{\sum_{i=1}^{m} t_i^2 - \left(\sum_{i=1}^{m} t_i\right)^2/m}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t}$$

where m = number pre-change points
$i = 1, 2, \ldots m$
$y = \sum_{i=1}^{m} y_i/m$
$\bar{t} = \sum_{i=1}^{m} t_i/m$

The difference between the estimated value and the obtained first post-change value, y_0, is used in the *t* test. Since $\sigma_u = \sqrt{\sigma^2\left[\frac{m+1}{m} + \frac{(t_0 - \bar{t})^2}{\sum_{i=1}^{m}(t_i - \bar{t})^2}\right]}$ and an estimate, $\hat{\sigma}^2$, of σ^2 is given by $\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{\alpha} - \hat{\beta}t_i)^2$, the computation formula for the *t* test of the difference between the obtained and predicted post-change point is given by

$$t = \frac{|y_0 - \hat{y}_0|}{\sqrt{\left(\frac{m+1}{m} + \frac{(t_0 - \bar{t})^2}{\sum_{i=1}^{m}(t_i - \bar{t})^2}\right)\left(\frac{\sum_{i=1}^{m}(y_i - \hat{\alpha} - \hat{\beta}t_i)^2}{m-2}\right)}}$$

The denominator is the standard error of the difference for t. The df for t is (m-2). The significance of the obtained t is evaluated by reference to a standard t table.

(b) Walker-Lev Tests. Walker and Lev (1953, pp. 390-395) provide the following tests of significance. These tests are useful in determining whether differences exist between the regression equations for pre- and post-change groups.

Test One.

This is a test of the hypothesis of common slope. The F ratio is given by $F = \frac{S_1}{S_2} \cdot \frac{(N-4)}{1}$

where S_1 = sum of squares of the common within groups regression estimates from the separate group regression estimates

$$\sum_{i=1}^{2}\sum_{j=1}^{N_i}\left[(\hat{\alpha}_i+\hat{\beta}_i t_{ij})-(\hat{\alpha}_{w_i}+\hat{\beta}_w t_{ij})\right]^2$$

S_2 = sum of squares of the obtained occasion values from the separate group regression estimates

$$\sum_{i=1}^{2}\sum_{j=1}^{N_i}\left[Y_{ij}-(\hat{\alpha}_i+\hat{\beta}_i t_{ij})\right]^2$$

N = total number of occasions in time series
i=1,2 [1=prechange group;2=postchange group]
$j = 1, 2, \ldots N_i$

Formulas for the common within groups slope and intercept are

$$\hat{\beta}_w = \frac{CTY_w}{CTT_w}$$

$$\hat{\alpha}_{w_i} = \bar{Y}_i - \hat{\beta}_w \bar{t}_i$$

Formulas for the separate group slopes and intercepts are

$$\hat{\beta}_i = \sum_{j=1}^{N_i} t_{ij} Y_{ij} - \left(\sum_{j=1}^{N_i} t_{ij}\right)\left(\sum_{j=1}^{N_i} Y_{ij}\right)/N_i$$

$$\hat{\alpha}_i = \bar{Y}_i - \hat{\beta}_i \bar{t}_i$$

The computational formulae for the numerator and denominator sum of squares for the F ratio are

$$S_1 = \sum_{i=1}^{2}\frac{CTY_i}{CTT_i} - \frac{C^2TY_w}{CTT_w}$$

$$S_2 = CYY_\omega - \sum_{i=1}^{2} \frac{C^2TY_i}{CTT_i}$$

where

$$CTY_i = \sum_{j=1}^{N_i} t_{ij}\, y_{ij} - \left[\frac{\sum_{j=1}^{N_i} t_{ij} \sum_{j=1}^{N_i} y_{ij}}{N_i} \right]$$

$$CTT_i = \sum_{j=1}^{N_i} t_{ij}^2 - \frac{\left(\sum_{j=1}^{N_i} t_{ij} \right)^2}{N_i}$$

$$CYY_i = \sum_{j=1}^{N_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{N_i} y_{ij} \right)^2}{N_i}$$

$$CTY_\omega = \sum_{i=1}^{2} CTY_i$$

$$CTT_\omega = \sum_{i=1}^{2} CTT_i$$

$$CYY_\omega = \sum_{i=1}^{2} CYY_i$$

The resulting F ratio is evaluated with 1 and N-4 degrees of freedom.

Test Two.

This is a test of the hypothesis that the slopes of the pre- and post-change groups are equal to zero when it has been established that $\hat{\beta}_i = \hat{\beta}_\omega$ (that is, F of Test One is not significant). The variance ratio is given by

$$F = \frac{C^2TY_\omega}{CTT_\omega \cdot CYY_\omega - C^2TY_\omega} \cdot \frac{N-4}{1}$$

and is evaluated with 1 and N-4 degrees of freedom.

Test Three.

This is a test of the hypothesis that a single regression line fits both the pre- and post-change groups. The F ratio is given by

$$F = \frac{S_b}{S_\omega} \cdot \frac{N-3}{1}$$

where S_b = sum of squares of common within groups regression estimates from the total series regression estimates

$$\sum_{i=1}^{2}\sum_{j=1}^{N_i}\left[(\hat{\alpha}_{w_i}+\hat{\beta}_w t_{ij})-(\hat{\alpha}_T+\hat{\beta}_T t_{ij})\right]^2$$

S_w = sum of squares of obtained occasion values from the common within groups regression estimates

$$\sum_{i=1}^{2}\sum_{j=1}^{N_i}\left[Y_{ij}-(\hat{\alpha}_{w_i}+\hat{\beta}_w t_{ij})\right]^2$$

Formulas for the total series slope and intercept are

$$\hat{\beta}_T=\frac{CTY_T}{CTT_T}$$

$$\hat{\alpha}_T=\bar{Y}-\hat{\beta}_T\bar{t}$$

The computational formulae for the numerator and denominator sum of squares for the *F* ratio are

$$S_b=CYY_B+\frac{C^2TY_w}{CTT_w}-\frac{C^2TY_T}{CTT_T}$$

$$S_w=S_1+S_2$$

where

$$CTY_T=\sum_{i=1}^{2}\sum_{j=1}^{N_i}t_{ij}Y_{ij}-\left[\left(\sum_{i=1}^{2}\sum_{j=1}^{N_i}t_{ij}\right)\left(\sum_{i=1}^{2}\sum_{j=1}^{N_i}Y_{ij}\right)\Big/N_i\right]$$

$$CYY_T=\sum_{i=1}^{2}\sum_{j=1}^{N_i}Y_{ij}^2-\left[\left(\sum_{i=1}^{2}\sum_{j=1}^{N_i}Y_{ij}\right)^2\Big/N\right]$$

$$CTT_T=\sum_{i=1}^{2}\sum_{j=1}^{N_i}t_{ij}^2-\left[\left(\sum_{i=1}^{2}\sum_{j=1}^{N_i}t_{ij}\right)^2\Big/N\right]$$

$$CYY_B=CYY_T-CYY_w$$

The resulting F ratio is evaluated with 1 and N-3 degrees of freedom.

(c) "Double Extrapolation" Technique. This test is concerned with the significance of the difference between two separate regression estimates of the y-value for time t_o which lies midway between the last pre-change point and the first post-change point. Reference to the Interrupted Time Series Design diagrammed on page 2 will indicate that this point lies midway between t_m and t_{m+1}. Assuming that the points are equally spaced in time, t_o is equal to $t_{m+.5}$.

The first regression estimate for y_o is obtained from the pre-change values; the second regression estimate for y_o, from the post-change values. The formulas for the two regression estimates are

$$\hat{\beta}_i = \frac{\sum_{j=1}^{N_i} t_{ij} y_{ij} - \left(\sum_{j=1}^{N_i} t_{ij}\right)\left(\sum_{j=1}^{N_i} y_{ij}\right) / N_i}{\sum_{j=1}^{N_i} t_{ij}^2 - \left(\sum_{j=1}^{N_i} t_{ij}\right)^2 / N_i}$$

$$\hat{\alpha}_i = \bar{y}_{ij} - \hat{\beta}_i \bar{t}_{ij}$$

where $i = 1,2$ [1 = prechange group; 2 = postchange group]
N_i = number occasions in group i
$j = 1,2,\ldots N_i$

Thus, the two estimates for y_o are given by $\hat{y}_{oi} = \hat{\alpha}_i + \hat{\beta}_i t_o$. The difference between these two estimates can be evaluated by the t-ratio $t = \frac{|\hat{y}_{o1} - \hat{y}_{o2}|}{S_D}$ and its significance determined by reference to a t table with $N_1 + N_2 - 4$ degrees of freedom. The standard error of the difference for t depends upon the variability of the pre- and post-change values, N_1 and N_2, the relation between t and y, and the distance of t_o from $\bar{t}_1$ and $\bar{t}_2$. The formula for S_D is given in Walker and Lev (p. 400) as

$$S_D = \sqrt{\frac{\left[\sum_{i=1}^{2}\left(CYY_i - \frac{C^2TY_i}{CTT_i}\right)\right]\left[\frac{N}{N_1N_2} + \sum_{i=1}^{2}\frac{(t_o - \bar{t}_i)^2}{CTT_i}\right]}{N_1 + N_2 + 4}}$$

where

$$CY{\cdot}Y_i = \sum_{j=1}^{N_i} Y_{ij}^2 - \left[\left(\sum_{j=1}^{N_i} Y_{ij}\right)^2 \Big/ N_i\right]$$

$$CTT_i = \sum_{j=1}^{N_i} t_{ij}^2 - \left[\left(\sum_{j=1}^{N_i} t_{ij}\right)^2 \Big/ N_i\right]$$

$$CTY_i = \sum_{j=1}^{N_i} t_{ij} Y_{ij} - \left[\left(\sum_{j=1}^{N_i} t_{ij} \sum_{j=1}^{N_i} Y_{ij}\right) \Big/ N_i\right]$$

The second purpose of the present study was to determine what effect the violation of the assumption of uncorrelated error would have on the three tests of significance. Do the tests become inappropriate when there is positive autocorrelation between points?

The scientist usually does not have knowledge of the degree of sequential dependency which exists in a given time series. The first serial correlation coefficient (of lag one) is used to determine whether the observations of a given time series can be regarded as consisting of independent error only (measures not correlated). The serial correlation coefficient tests the hypothesis that the order of dependence in the time series is zero against the alternative that it is one. Serial correlation coefficients of lags higher than one can also be considered. (If the serial correlation is that of a variable with itself, it may be referred to as an auto-correlation coefficient.)

A serial correlation coefficient of lag one, r_1, is obtained by pairing observations one time unit apart; that is, the first observation is paired with the second, the second with the third, etc., throughout the entire time series until the last observation is reached. The product-moment correlation is computed using the resulting pairs of observations. In a similar manner, the second serial correlation coefficient, r_2, is obtained by pairing successive observations two time units apart and r_m is obtained by pairing observations m units apart. The basic formula for r_a is given below:

$$r_a = \frac{n \sum_{i=1}^{n} y_i y_{i+a} - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_{i+a}}{\sqrt{\left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]\left[n \sum_{i=1}^{n} y_{i+a}^2 - \left(\sum_{i=1}^{n} y_{i+a}\right)^2\right]}}$$

where N = total number of occasions in the time series
$n = N-a$
$i = 1,2,\ldots n$

The model of autocorrelation used in the present study is basically one of proximity, namely, it was assumed that measurements made closer together in time would be more strongly related than measurements made further apart in time. This means that in instances where a significant autocorrelation r_a exists, increasingly larger autocorrelations should exist for r_{a-1}, r_{a-2}, ... and r_1, respectively.

Since the slope of the line contributes to serial dependency between points, interest in the present study is in the autocorrelations of departures from the line of best fit for the total series. However, because true effects increase proximal autocorrelation, differences from the separate pre- and post-change regression lines give better estimates of existing autocorrelations in instances of true effects. So as not to penalize true effects in the data version of the program (a computer program which performs the tests of significance on input data of interrupted time series form is presented in Sween and Campbell, 1965), autocorrelation coefficients based on departures from the separate regression lines are also computed. Although only departures from the line of best fit for the total series were used in the present study, in further work autocorrelations based on differences from separately fitted regression lines will also be utilized to increase comparability of the Monte Carlo results with actual experimental data.

METHOD

The procedure was basically that of generating a large number of time series having specified characteristics and performing the tests of significance on each generated time series. In this way, distributions of _t_'s and/or _F_'s for the three tests of significance were obtained. The distributions were then examined to determine how satisfactory each of the three tests of significance were in terms of the risk (α) of rejecting the null hypothesis (the experimental change _X_ had no effect) when it was true.

The time series were constructed so that the hypothesis of no effect was true. Normal random error was added to each "true" point to produce a time series of _N_ observed values such that $y_{i(observed)} = y_{i(true)} + error_i$ for i = 1,...N. Two general types of error, independent error and/or correlated error, could be added to the "true" line values. When the null hypothesis is true and the assumptions for use of the tests of significance are met, the theoretical values from the _t_ and _F_ tables should be exceeded by chance only 1% and 5% of the time. Thus, when only independent error is added, the discrepancies between the theoretical values and the obtained percent of significant _t_'s and _F_'s should indicate how suitable the test of significance is in the interrupted time series situation. In addition, the obtained percentage of t's and F's which exceed the theoretical values when sequential dependency between points is built in should indicate how the violation of the assumption of independence of error may further restrict the usefulness of the tests.

Each generated time series could be varied with respect to (a) the number of pre- and post-change occasion values, (b) the slope of the "true" line, (c) the degree of autocorrelation between points, and (d) the total error variance about the true line values. In the present study the following combinations of pre- and post-change occasion values, "true" line slopes, autocorrelated error, and error variance were used:

(a) Pre- and post- change sample sizes of 10, 20, and 100 were used to yield time series with the following number of total occasion values:

Total Occasions, N = 20 (pre = 10; post = 10)
Total Occasions, N = 40 (pre = 20; post = 20)
Total Occasions, N = 200 (pre =100; post =100)

In the presentation of the results these three conditions of sampling are referred to in terms of the total occasions in the series as N = 20, N = 40, and N = 200. The degrees of freedom and critical values for the Mood test of significance are, however, based only upon the pre-change points of 10, 20, and 100. For the Walker-Lev Test 3 and Double Extrapolation Technique the total series points of 20, 40, and 200 are used.

(b) The slope of the true line was specified as 0.0, 1.0, and 20.0.

(c) The degree of autocorrelated error was specified as zero (independent error only), one, two, and three (correlated error for measurements one, two, and three time lags apart).

(d) Total error variance about the true line was specified at 1.00 and 5.00 and normal random error was drawn from Gaussian distributions of zero mean and appropriate standard deviations to yield equal error variance about the true line for all degrees of autocorrelated error.

For the total error variance specified as 1.00, the standard deviations of the normal distributions from which the normal random errors were drawn would be

Supplementary Note:

Since the research reported herein was done, a highly relevant test of significance has been reported:

G. E. P. Box and George C. Tiao, "A change in level of a non-stationary time series," Biometrika, 1965, 52, 181-192.

In particular, the Box and Tiao approach avoids the assumption of linearity, in exchange for other, probably more reasonable, assumptions. What is needed is a computer program for the exact distribution computation by numerical methods for their test when δ and γ_o are unknown (pp. 189-191), plus a testing of the formula against null models such as those used here, plus a testing of our linear formulas against the null data generated according to the Box and Tiao assumptions.

1.00 (unique error only)
.58 (unique plus error of lag 1)
.50 (unique plus error of lag 2)
.45 (unique plus error of lag 3)

For a total error variance specified as 5.00, the standard deviations of the normal distributions from which the normal random errors were drawn would be

2.24 (unique error only)
1.29 (unique plus error of lag 1)
1.12 (unique plus error of lag 2)
1.00 (unique plus error of lag 3)

One thousand sets of time series were generated for each type of autocorrelated error in various combinations with the options of sample size, true line slope, and total error variance listed in a, b, and d above. The following tests of significance were performed on each of the 1000 sets of generated data: Mood test, Walker-Lev Test 3, Double Extrapolation technique. [The Walker-Lev Tests 1 and 2 and a test proposed by Clayton and described by Campbell (1963, p. 225) were performed on a smaller sample of 100 sets of generated time series.]^1

For each test of significance, the percent of t's or F's which exceeded the theoretical 1% and 5% values was determined. The complete sequence of operations was performed internally on the IBM 709 computer system programmed to perform the necessary operations. The basic steps in the computation procedures are summarized below:

(1) The "true" line pre- and post-X occasion values were determined on the basis of desired slope. Normal random errors yielding specified autocorrelation were added to the "true" line points to form the set of time series data. (A binary subprogram from the Vogelback computing center (NU-0044) was used for the generation of normal random numbers. The mean of the Gaussian distribution approximates zero; the standard deviation was specified as described in (d) above). The program generated 1000 sets of time series for each type of error fluctuation specified.

(2) F and t values for the Mood, Walker-Lev 3, and Double Extrapolation tests of significance and serial correlation coefficients r(1), r(2),

r(3), and r(4) were computed for each of the 1000 sets of generated time series data. A data plot of the time series could also be obtained for each set.

(3) The 1000 t and F values for the Mood, Walker-Lev 3, and Double Extrapolation tests were sorted on the basis of magnitude. The ordered t and F values and/or the percents of *t*'s and *F*'s above the tabled critical values were printed out. The correlations between the t and F values and the autocorrelation coefficients were determined.

RESULTS

The results of the present study indicated that the three main tests of significance (Mood test, Walker-Lev Test 3, and Double Extrapolation Technique) are appropriate for use on data of interrupted time series form. However, the results also indicated that when statistical dependency between measurements exists use of these tests with the tabled critical values will yield significant results by chance alone *more than* the expected one percent and five percent of the time. This was particularly true for the Walker-Lev Test 3 and the Double Extrapolation Technique.

These results are summarized in Table 1 where the percent of *F*'s and *t*'s above the tabled 1% and 5% critical values are given for the three statistical tests of significance. These percent values were obtained from the

---

Insert Table 1 about here

number of instances of significance in 1000 sets of generated Monte Carlo time series. Four degrees of dependency between points and total numbers of time series occasion values of 20, 40, and 200 are represented in Table 1 (for a total N of 200, only the independent error and the lag three correlated error Monte Carlo generations were available). As indicated in Table 1, when no significant autocorrelation exists between points, that is, the errors are independent, the alpha values are approximately the theoretical

Table 1

Percent of F's and t's Above One and Five Percent Critical Values as a Function of Degree of Correlated Error

		One Percent				Five Percent			
		Independent Error	Correlated Error			Independent Error	Correlated Error		
Test of Significance	Total Occasions in Time Series		Lag 1	Lag 2	Lag 3		Lag 1	Lag 2	Lag 3
Mood Test	N = 20	1.10*	1.20	2.00	2.10	5.30	5.40	7.00	9.50
	N = 40	1.00	1.50	1.90	1.70	5.80	5.50	5.90	9.10
	N = 200	.90	--	--	.90	5.20	--	--	5.80
Walker-Lev Test 3	N = 20	1.10	2.80	5.90	9.20	5.60	9.60	17.30	19.80
	N = 40	1.40	3.30	8.50	13.20	5.20	11.50	20.20	25.30
	N = 200	1.40	--	--	17.50	5.90	--	--	28.80
Double Extrapolation	N = 20	1.10	2.90	7.40	11.10	5.30	11.10	19.20	23.40
	N = 40	1.50	3.60	9.50	15.10	5.00	11.60	20.60	27.10
	N = 200	1.40	--	--	17.70	5.70	--	--	28.40

*Each percent based on number of significant instances in 1000 sets of generated time series of true line slope = 1.00 and total error variance = 5.00.

values expected. When the assumption of independence is not met (correlated error), the percent of significant time series exceeds the expected one percent and five percent values. The percent of significant instances increases with increasing autocorrelation between points, particularly for the Walker-Lev Test 3 and Double Extrapolation Technique.

The slope of the true line and the total error variance about the true line had no effect on the percent of F's and t's above the tabled critical values. The percents of false-positives were similar for true line slopes of 0, 1, and 20 and for total error variances of 1.00 and 5.00. These results were obtained for all three tests of significance.

Although the true line slope and total error variance had no effect on the number of false-positives, the total number of occasions in the time series did produce an effect. As can be observed in Table 1, the percent of significant instances tends to vary with the total N of the series. This is particularly evident for the Walker-Lev Test 3 and Double Extrapolation Technique performed on time series with correlated error for three lags. The greater error in regression estimates with smaller N's is most likely the crucial factor in this effect.

The effect of the total number of occasion values in the time series was also evident in departures of the obtained average autocorrelation coefficients from their expected values. The expected values ρ_a of r_a is given by $\rho_a = \mathrm{cov}\{Y_i Y_{i+a}\} / \sqrt{\sigma^2_{Y_i} \sigma^2_{Y_{i+a}}}$ and the standard error of r_a, by $\sigma_{r_a} = 1 - \rho_a^2 / \sqrt{N_s - 1}$ where N_s refers to the sample size of 1000 in the present study. When expected values were compared with the obtained averages (over 1000 sets of generated time series) the obtained averages were less than the expected values for the smaller N's of 20 and 40. The obtained averages approached the expected values for N of 200. Table 2 indicates this relationship between expected value and obtained averages for the first autocorrelation coefficient r_1. Similar results were obtained for the autocorrelation coefficients r_2, r_3, r_4.

Insert Table 2 about here

On the basis of the Monte Carlo findings of elevated alpha values for time series exhibiting proximally correlated error, we present in Table 3 new critical values for the three tests of significance. These new critical values approximate more closely the expected one and five percent significance levels. The new critical values were obtained by the method of basic linear interpolation from the IBM printout giving the "Percent of *Z*'s and *t*'s in intervals of specific widths" for each test of significance. The values are based on Monte Carlo generations of 1000 sets of time series for each type of error fluctuation. As shown in Table 3, when time series observations are not correlated (assumption of independence met) the critical values which yield significant results one and five percent of the time by chance alone are, in general, similar to the tabled critical values for all three statistical tests of significance. However, when a significant autocorrelation exists between points, that is, when a given observation is dependent upon preceding observations, the critical values yielding significant results one and five percent of the time are similar to the tabled values for the Mood test only. For both the Walker-Lev Test 3 and Double Extrapolation Technique, the critical values increase with increasing autocorrelation.

Insert Table 3 about here

The new critical values (Table 3) are plotted in Figures 1-6 as a

Table 2

Means and Standard Deviations of the First Autocorrelation Coefficient r_1

		Obtained Values			Expected Values
Error Type		N = 20	N = 40	N = 200	
Independent	$\bar{x}$	-.11	-.05	-.01	.00
	SD	.21	.16	.07	.03
Correlated Lag 1	$\bar{x}$	.18	.26	--	.33
	SD	.20	.13	--	.03
Correlated Lag 2	$\bar{x}$	.31	.41	--	.50
	SD	.22	.14	--	.02
Correlated Lag 3	$\bar{x}$	.37	.49	.58	.60
	SD	.24	.14	.06	.02

Table 3

Critical Values Yielding One and Five Percent Significance Levels

		One Percent			Five Percent		
Test of Significance		N = 20*	N = 40	N = 200	N = 20	N = 40	N = 200
Mood Test	Degrees of Freedom for t	df=8	df=18	df=98	df=8	df=18	df=98
	Tabled Critical Values of t	3.36	2.88	2.60	2.31	2.10	1.99
	Independent Error	3.45	2.94	2.59	2.36	2.13	2.02
	Correlated Error-Lag 1	3.32	3.08	----	2.35	2.18	----
	Correlated Error-Lag 2	3.74	3.28	----	2.51	2.22	----
	Correlated Error-Lag 3	3.62	3.30	2.59	2.77	2.42	2.08
Walker-Lev Test 3	Degrees of Freedom for F	df=1,17	df=1,37	df=1,197	df=1,17	df=1,37	df=1,197
	Tabled Critical Values of F	8.40	7.35	6.76	4.45	4.10	3.89
	Independent Error	8.58	8.58	8.25	4.72	4.20	4.29
	Correlated Error-Lag 1	10.68	11.06	----	6.22	5.96	----
	Correlated Error-Lag 2	17.25	18.12	----	9.10	9.25	----
	Correlated Error-Lag 3	23.75	24.75	24.75	11.25	13.44	14.04
Double Extrapolation	Degrees of Freedom for t	df=16	df=36	df=196	df=16	df=36	df=196
	Tabled Critical Values of t	2.92	2.70	2.58	2.12	2.03	1.96
	Independent Error	3.03	2.90	2.85	2.14	2.04	2.05
	Correlated Error-Lag 1	3.52	3.36	----	2.56	2.49	----
	Correlated Error-Lag 2	4.64	4.46	----	3.24	3.16	----
	Correlated Error-Lag 3	5.14	4.98	4.95	3.57	3.75	3.73

*N refers to the total number of occasions in the generated time series. Degrees of freedom and critical values for the Mood test are based on 10, 20, and 100 pre-change points. The total series points of 20, 40, and 200 are used for the Walker-Lev Test 3 and Double Extrapolation Technique.

function of the obtained autocorrelation coefficient. As discussed above, the obtained average r_1 is less than the expected value due to greater error in regression estimates with small N. Figures 1-6 may be used to obtain approximate one percent and five percent critical values when correlated error exists between points. The new critical values are found by interpolating on the abscissa using the obtained first autocorrelation coefficient and on the face of the graph using N. For the Mood test N refers to the number of pre-change points; for the Walker-Lev Test 3 and Double Extrapolation Technique N refers to the total number of occasions in the time series. The new critical value is read from the ordinate.

Insert Figures 1-6 about here

Some Utilizations

The tests of significance which we are offering are being applied in situations in which the visual impression is the usual basis of interpretation. In some sense, the tests are designed to imitate such judgments, making more precise and rationalizing the criteria involved. Thus, the more variable the pre-change points, the larger the cross-treatment change must be to "appear" significant. And holding this constant at a small variability, the longer the sampling of pre- and post-change observations, the more confident we are that a given trans-treatment shift is truly exceptional, is more than a coincidence. These ingredients feature prominently in the tests we have examined.

In the present study, we have examined only linear hypotheses. Many times a significant effect in terms of linear hypotheses will appear upon graphic inspection to be a homogenous curvilinear process with no discontinuity

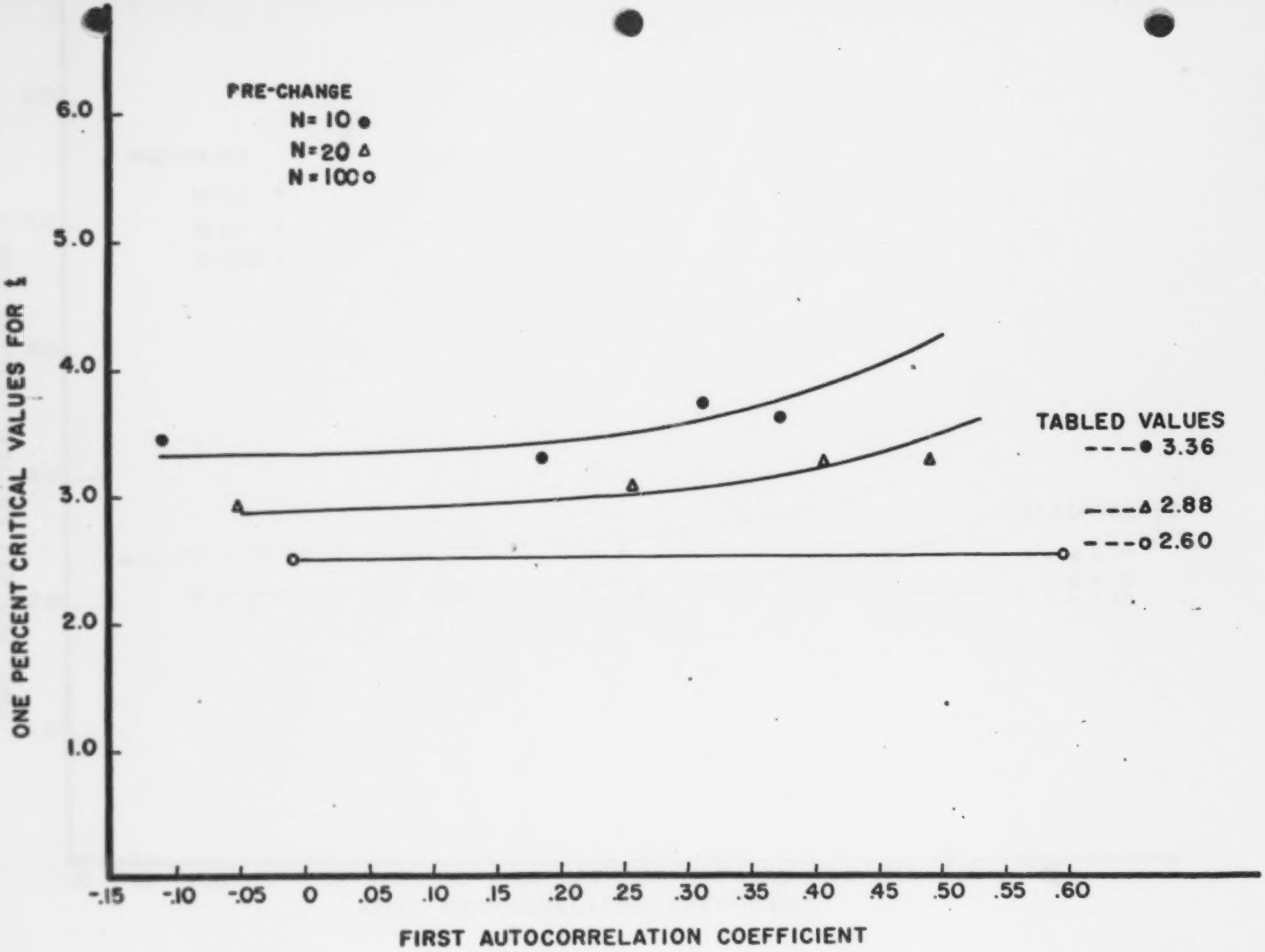


Fig. 1. Mood test: Obtained one percent critical values of t as a function of the first autocorrelation coefficient.

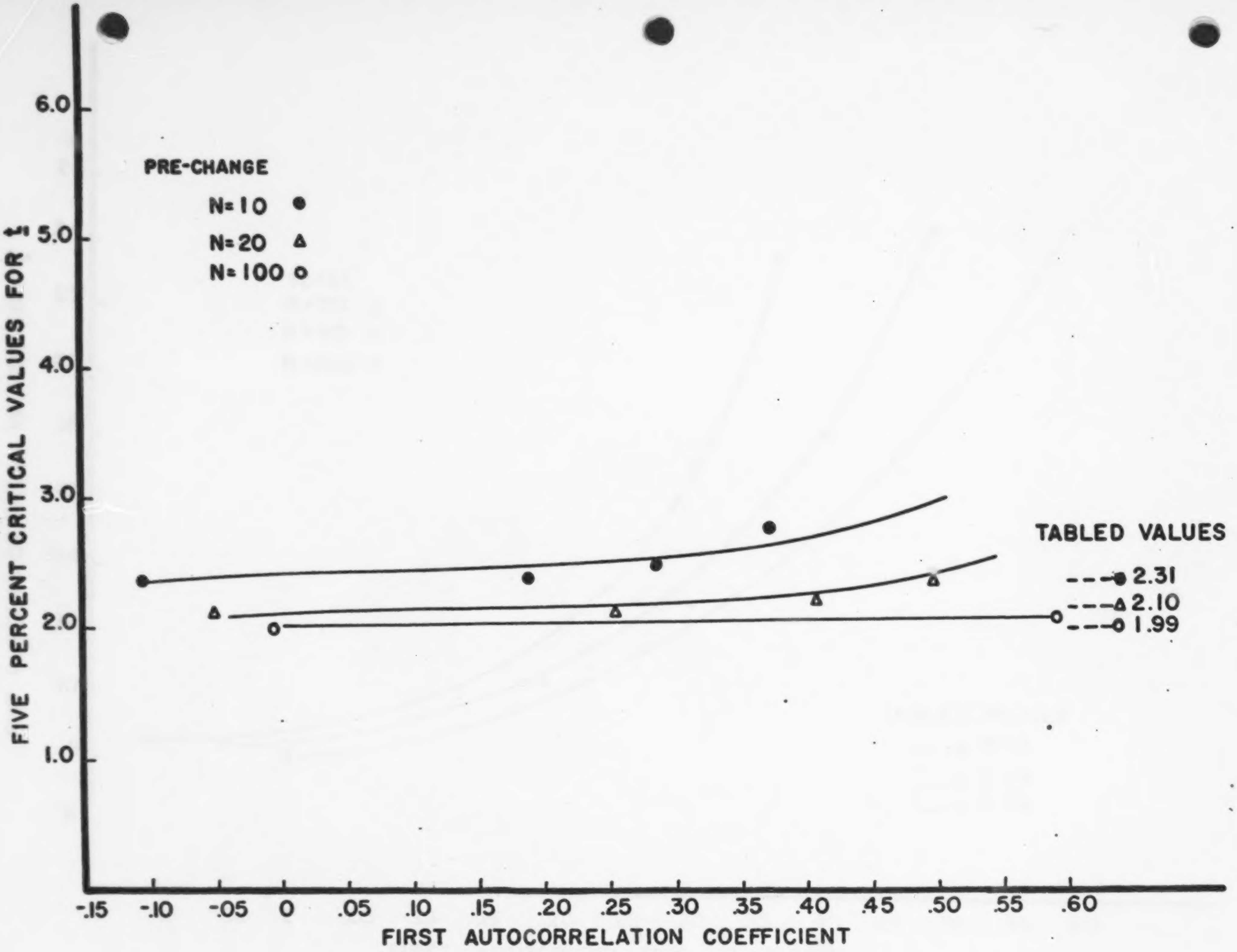


Fig. 2. Mood test: Obtained five percent critical values of t as a function of the first autocorrelation coefficient.

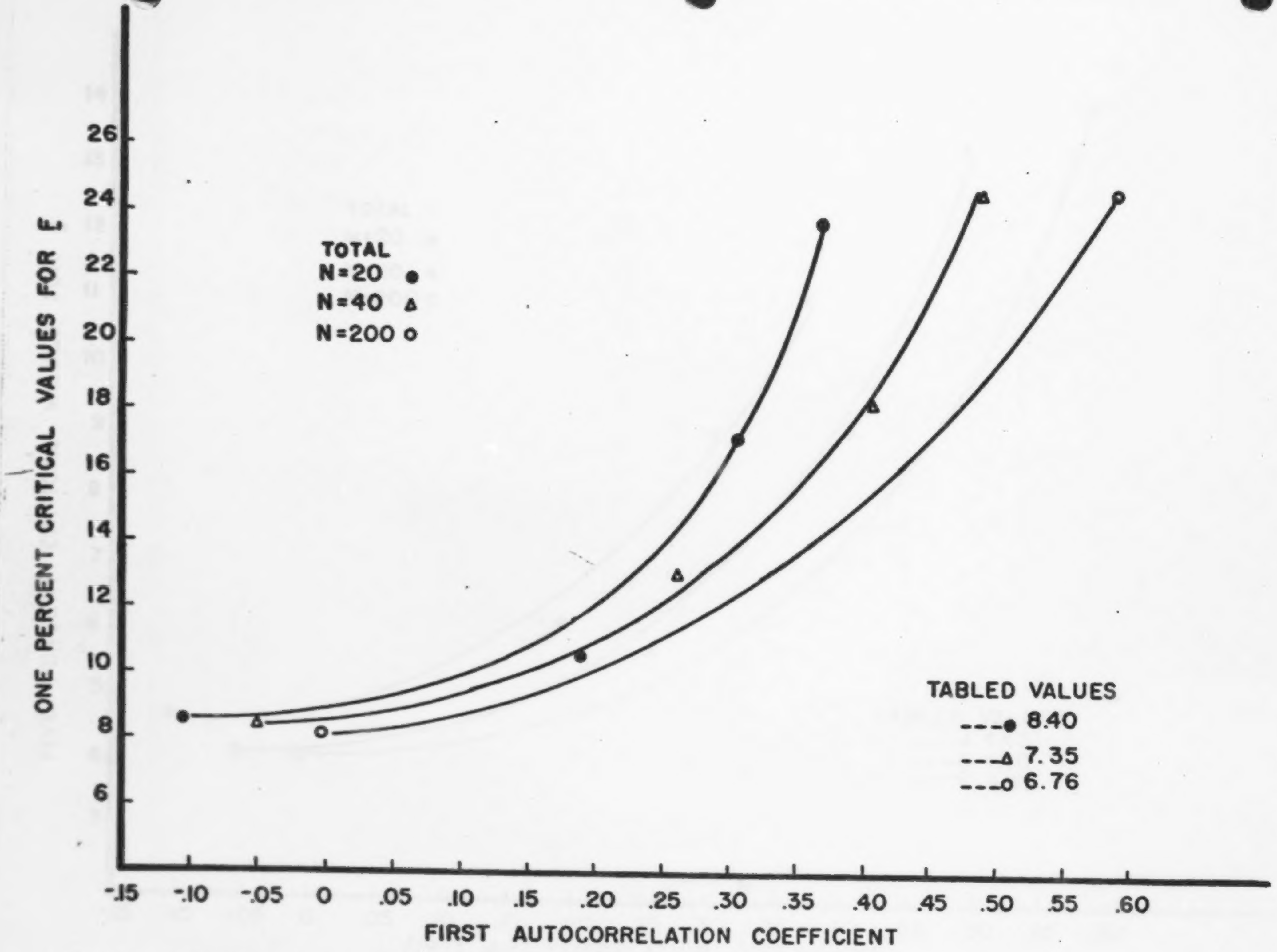


Fig. 3. Walker-Lev Test 3: Obtained one percent critical values of F as a function of the first autocorrelation coefficient.

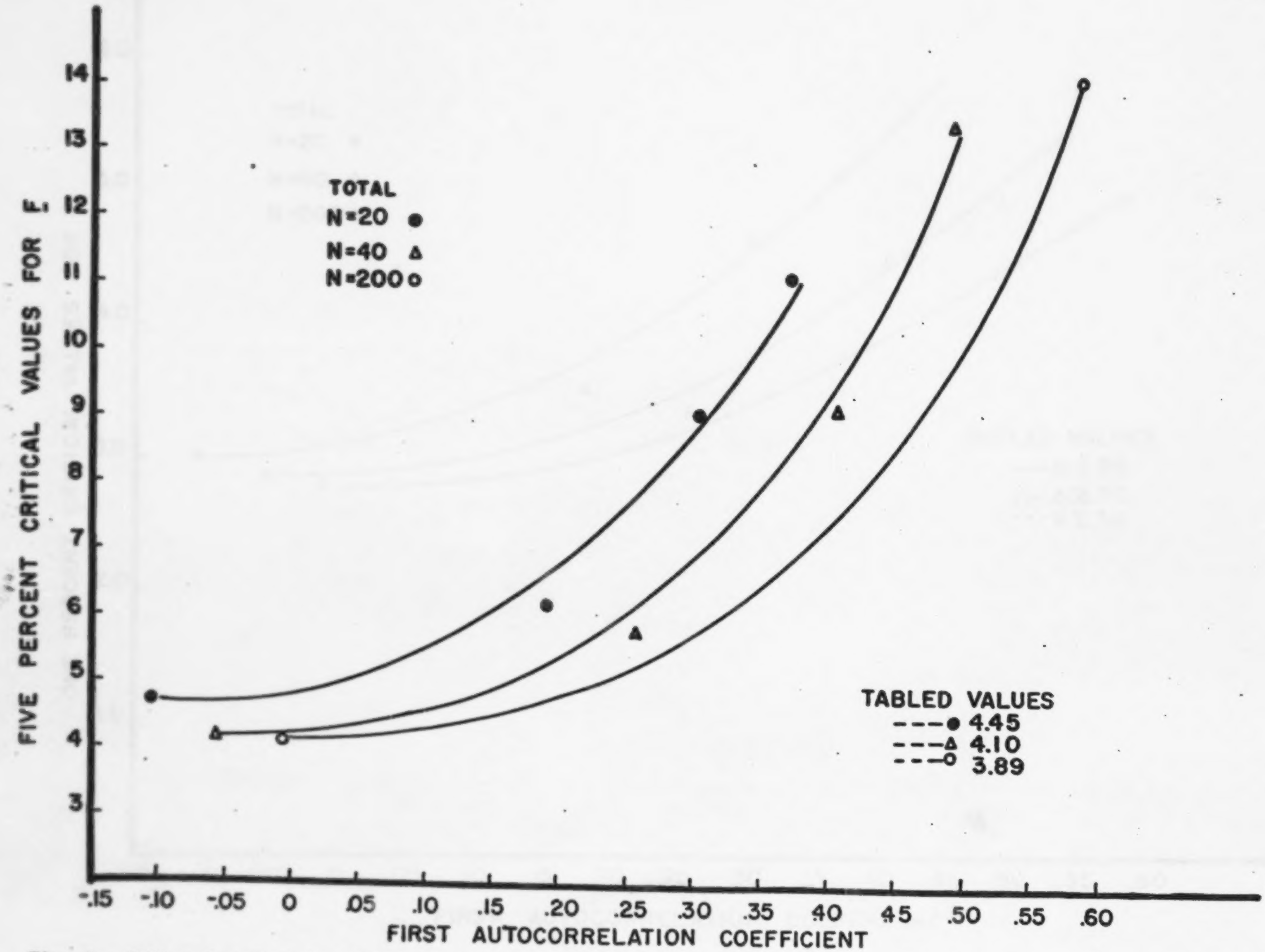


Fig. 4. Walker-Lev Test 3: Obtained five percent critical values of F as a function of the first autocorrelation coefficient.

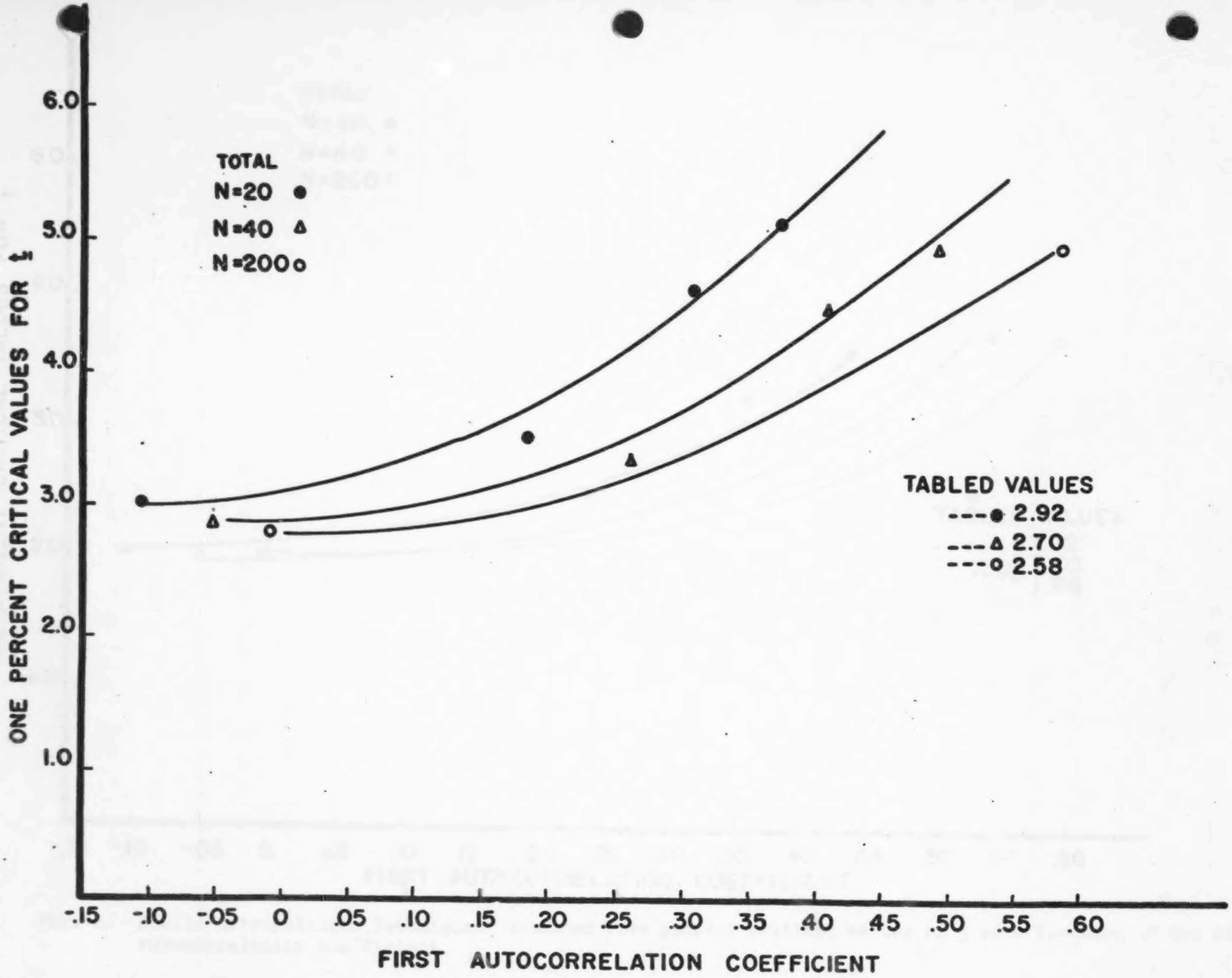


Fig. 5. Double Extrapolation Technique: Obtained one percent critical values of t as a function of the first autocorrelation coefficient.

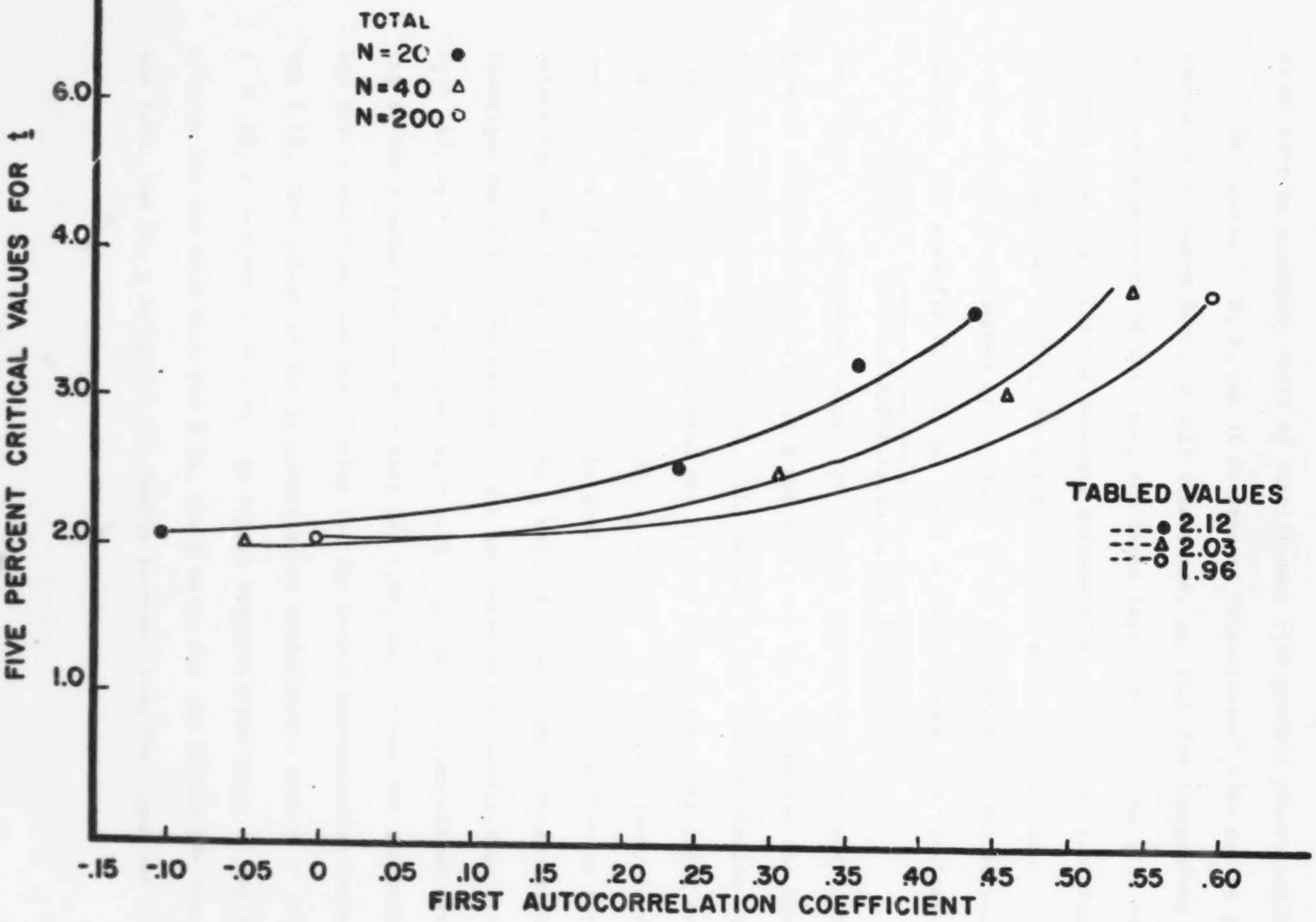


Fig. 6. Double Extrapolation Technique: Obtained five percent critical values of t as a function of the first autocorrelation coefficient.

at the treatment point. Usually one will not have sufficient degrees of freedom to test curvilinear hypotheses. For these and other reasons, it seems best to accompany tests of significance with graphic presentations.

In Figures 7, 8, 9, and 10 we present "significant" time series generated by the Monte Carlo of null conditions, one each for independent error and correlated error of one, two, and three lags. These graphs give some indication of the effect of proximal autocorrelations. In all four figures, the true line slope is 1.0, the total error variance about the true line is 5.00, and the total number of occasions is 40 (20 pre-change and 20 post-change). The specific data on the tests of significance is as follows:

Insert Figures 7, 8, 9, and 10 about here

For Figure 7 (independent error) the *t* value for the Mood Test was 2.55, the *F* value for the Walker-Lev Test 3 was 6.11, and the *t* value for the Double Extrapolation Technique was 2.45. The values of the autocorrelation coefficients based on departures from the line of best fit for the total series were $r_1 = .07$, $r_2 = -.24$, $r_3 = -.08$, $r_4 = -.07$. In Figure 8 (correlated error of lag 1) the *t* value for the Mood test was 2.34, the *F* value for the Walker-Lev Test 3 was 6.13, and the *t* value for the Double Extrapolation Technique was 2.45. The values of the autocorrelation coefficients were $r_1 = .45$, $r_2 = -.22$, $r_3 = -.44$, $r_4 = -.26$. In Figure 9 (correlated error of lag 2) the *t* value for the Mood test was 3.86, the *F* value for the Walker-Lev Test 3 was 4.66, and the *t* value for the Double Extrapolation Technique was 2.16. The values of the autocorrelation coefficients were $r = .51$, $r = .25$, $r = -.06$, $r = -.15$. In Figure 10(correlated error of lag 3) the *t* value for the Mood test was 3.54, the *F* value for the Walker-Lev Test 3 was 7.45, and the *t* value for the Double Extrapolation Technique was 2.88.

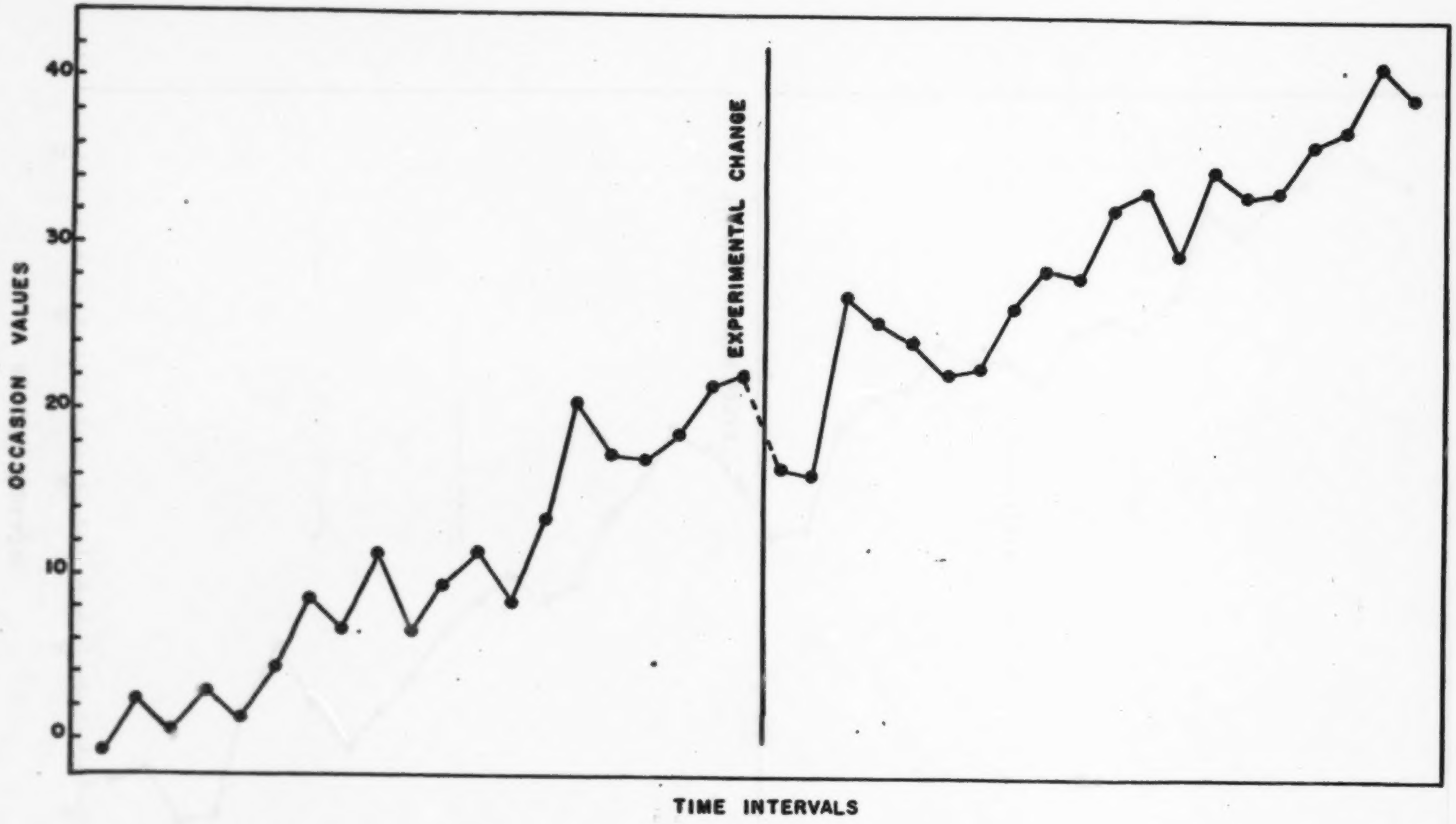


Fig. 7. A significant Monte Carlo generated time series for independent error.

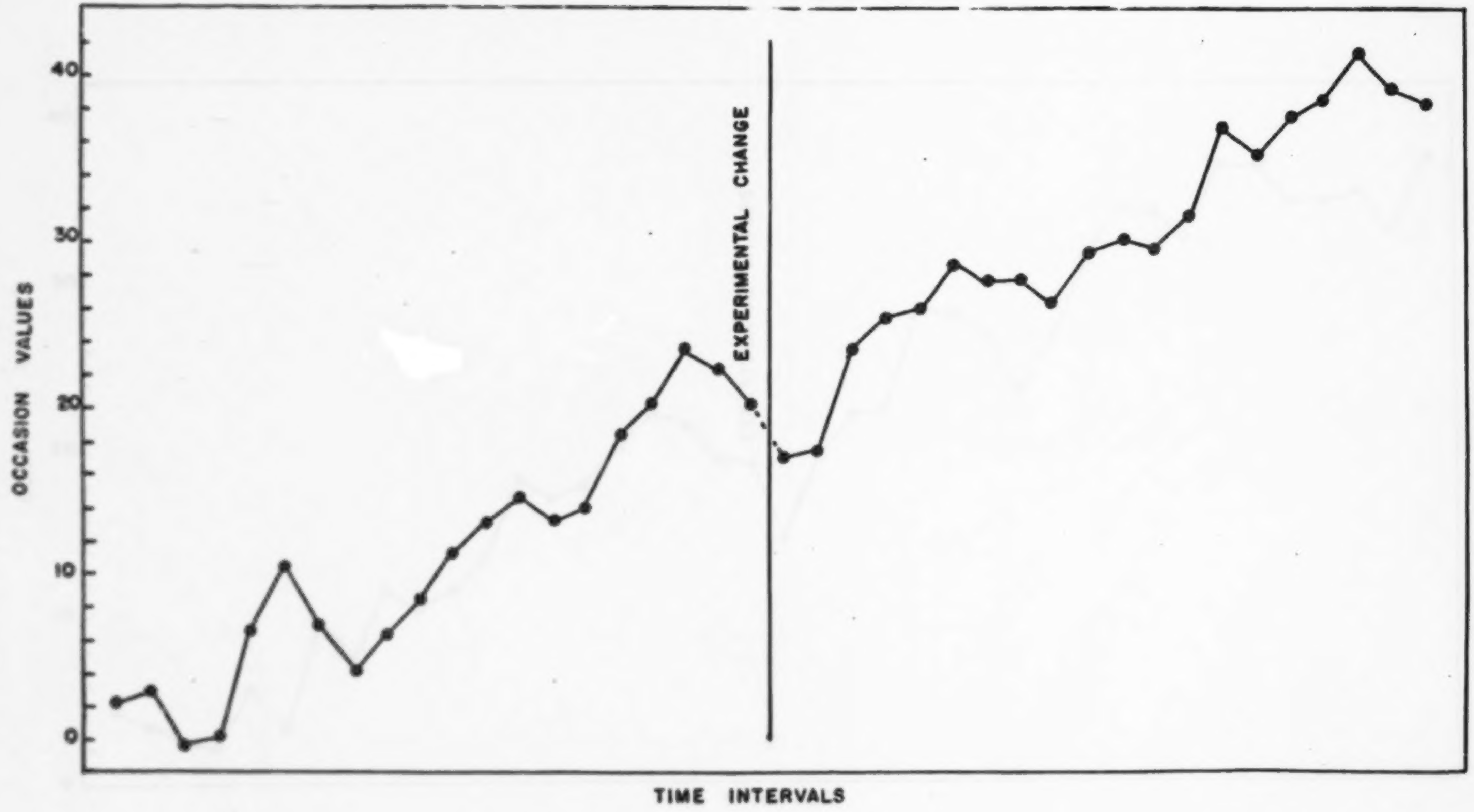


Fig. 8. A significant Monte Carlo generated time series for correlated error of one lag.

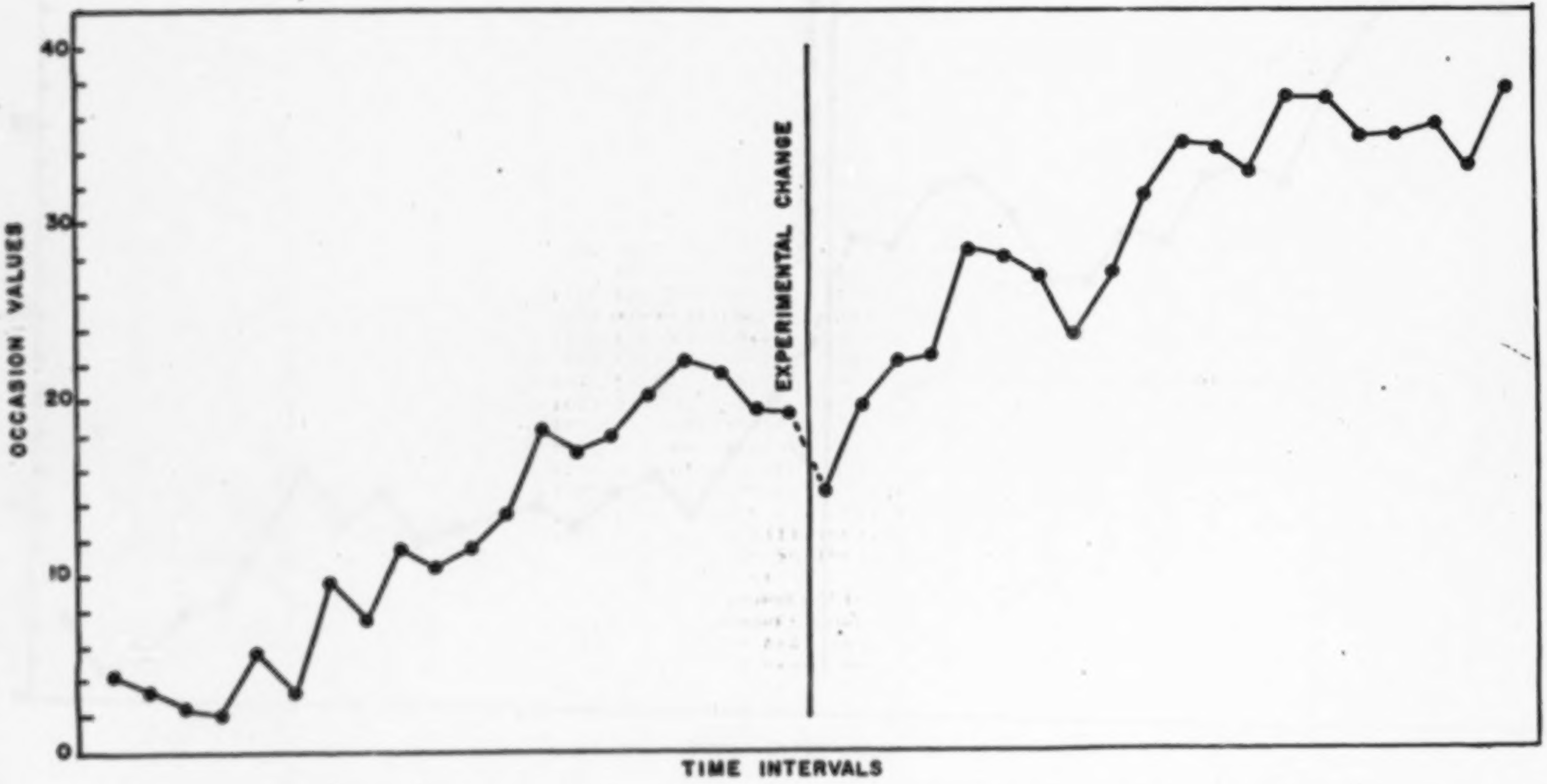


Fig. 9. A significant Monte Carlo generated time series for correlated error of two lags.

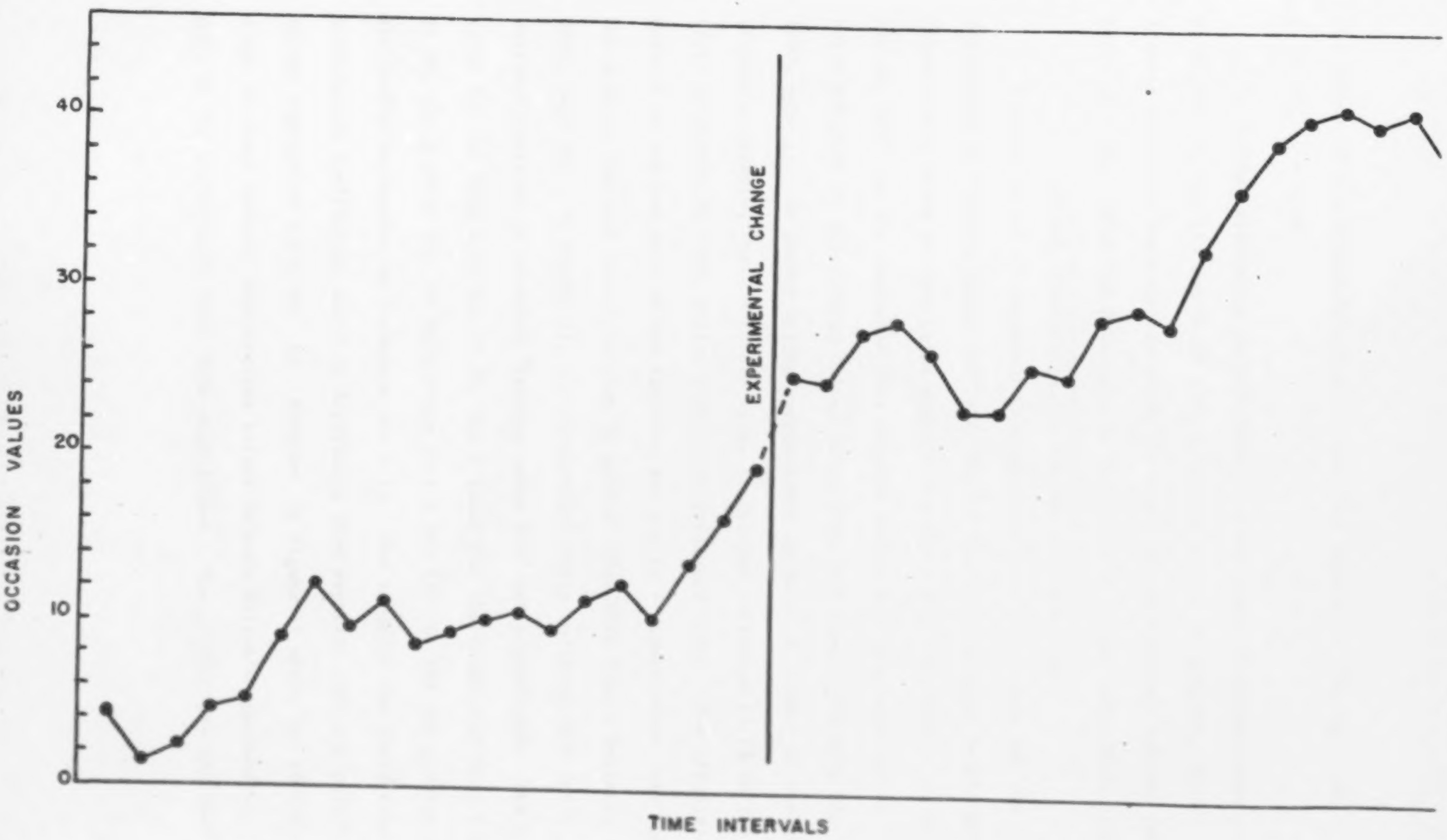


Fig. 10. A significant Monte Carlo generated time series for correlated error of three lags.

The values of the autocorrelation coefficients were $r_1 = .76$, $r_2 = .50$, $r_3 = .20$, $r_4 = -.14$.

To further illustrate utilizations of these tests of significance, in Figure 11, 12, 13, and 14 we present actual data. In general, these figures represent instances in which the tests of significance confirm judgments of effect which had been made on the basis of visual inspection alone.

Insert Figures 11, 12, 13, and 14 about here

Figures 11 and 12 represent Chicago crime rate statistics for the categories of "Larceny under $50" and "Murder and non-negligent manslaughter," respectively (from Uniform Crime Reports for the U.S., 1942-1962). On February 22, 1960, it was announced that Orlando Wilson had been selected as Superintendent of the Chicago Police Force (New York Times, February 23, 1960, page 1). He became acting commissioner on March 2, 1960. It was subsequently reported that "Recorded crime in Chicago increased 83.7% in the first 10 months of 1960, police statistics indicated today. The officials refused to say how much of the increase was due to a higher crime rate and how much to improved record keeping by police" (New York Times, December 17, 1960, page 34). In Figure 11, all statistical tests indicate that the observed increase in recorded "Larceny under $50" was significant. The t value for the Mood test was 12.54, the F value for the Walker-Lev Test 1 was 51.59, the F value for the Walker-Lev Test 3 was 157.76, and the t value for the Double Extrapolation Technique was 9.12. The value of the first autocorrelation coefficient based on departures from separate pre- and post-change regression lines was .43. However, in Figure 12 where the effect seems to start several observations before Orlando Wilson's appointment, none of the statistical tests were significant. The t value for the Mood

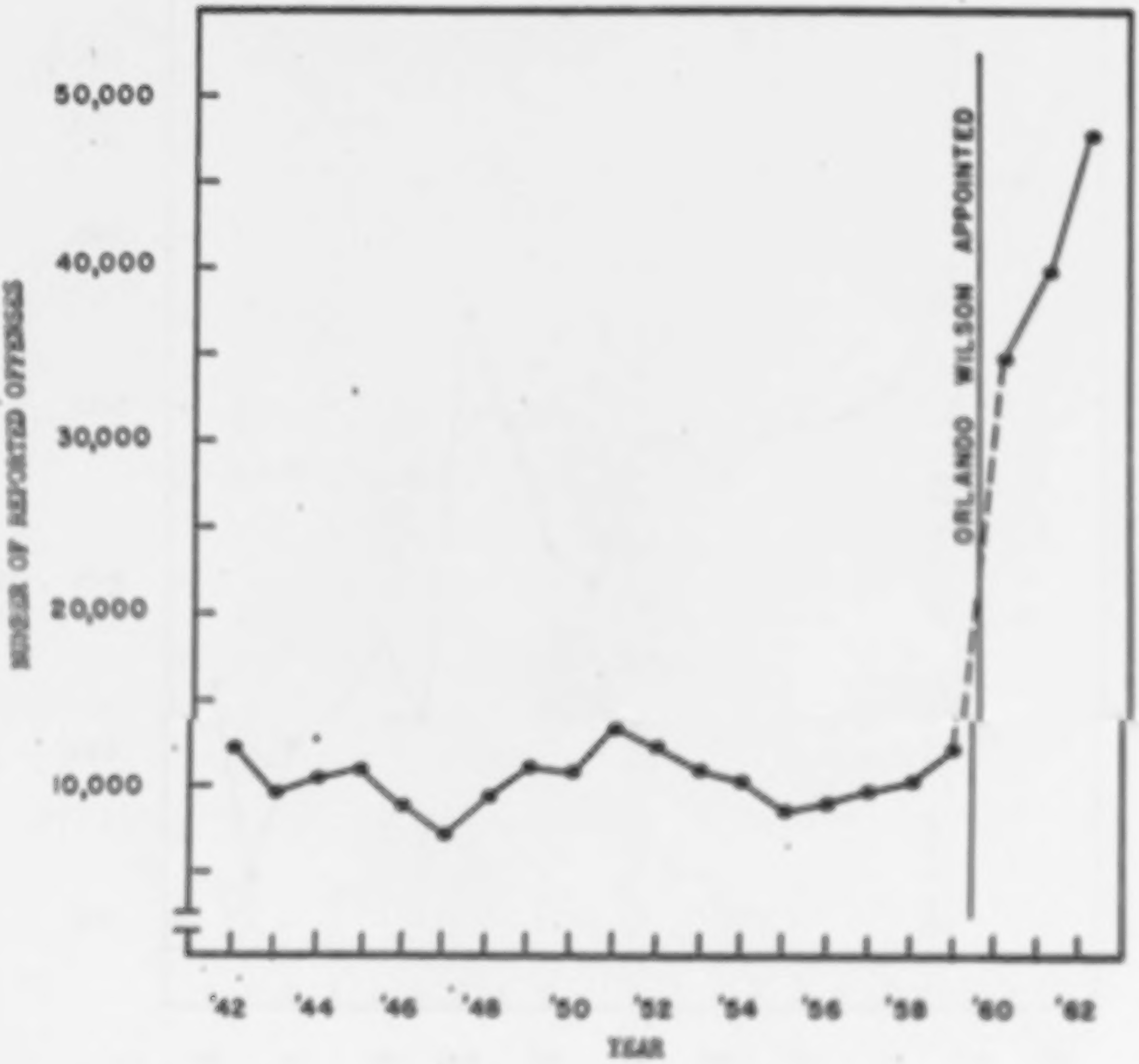


Fig. 11. Number of reported larcenies under $50. Chicago, Illinois. 1942-1962 (from *Uniform Crime Reports for the United States*, 1942-62).

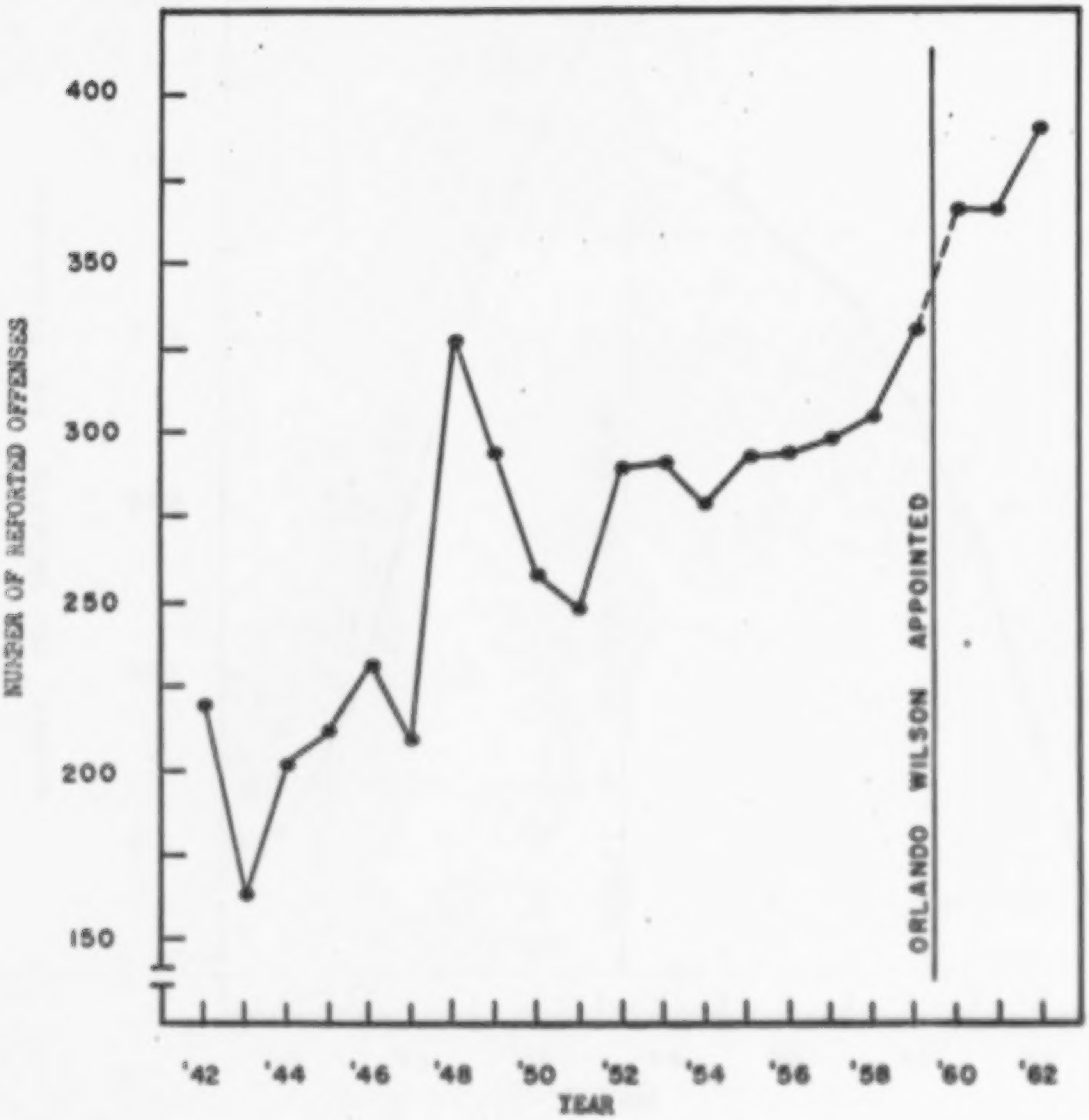


Fig. 12. Number of reported murders and non-negligent manslaughters. Chicago, Illinois. 1942-1962 (from Uniform Crime Reports for the United States, 1942-62).

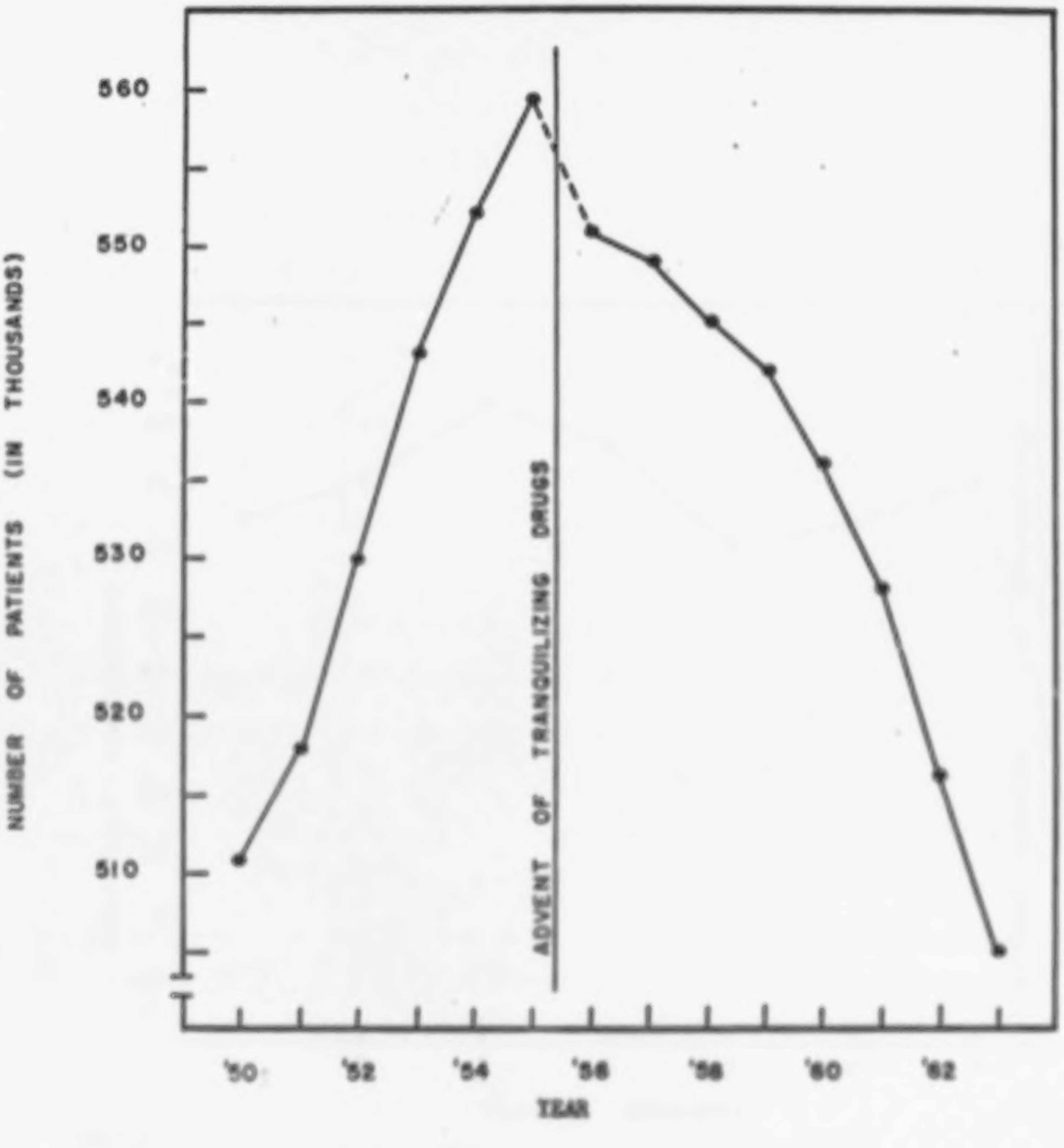


Fig. 13. Number of hospitalized mental patients in the United States before and after the advent of tranquilizing drugs (from *Britannica Book of the Year, 1964*).

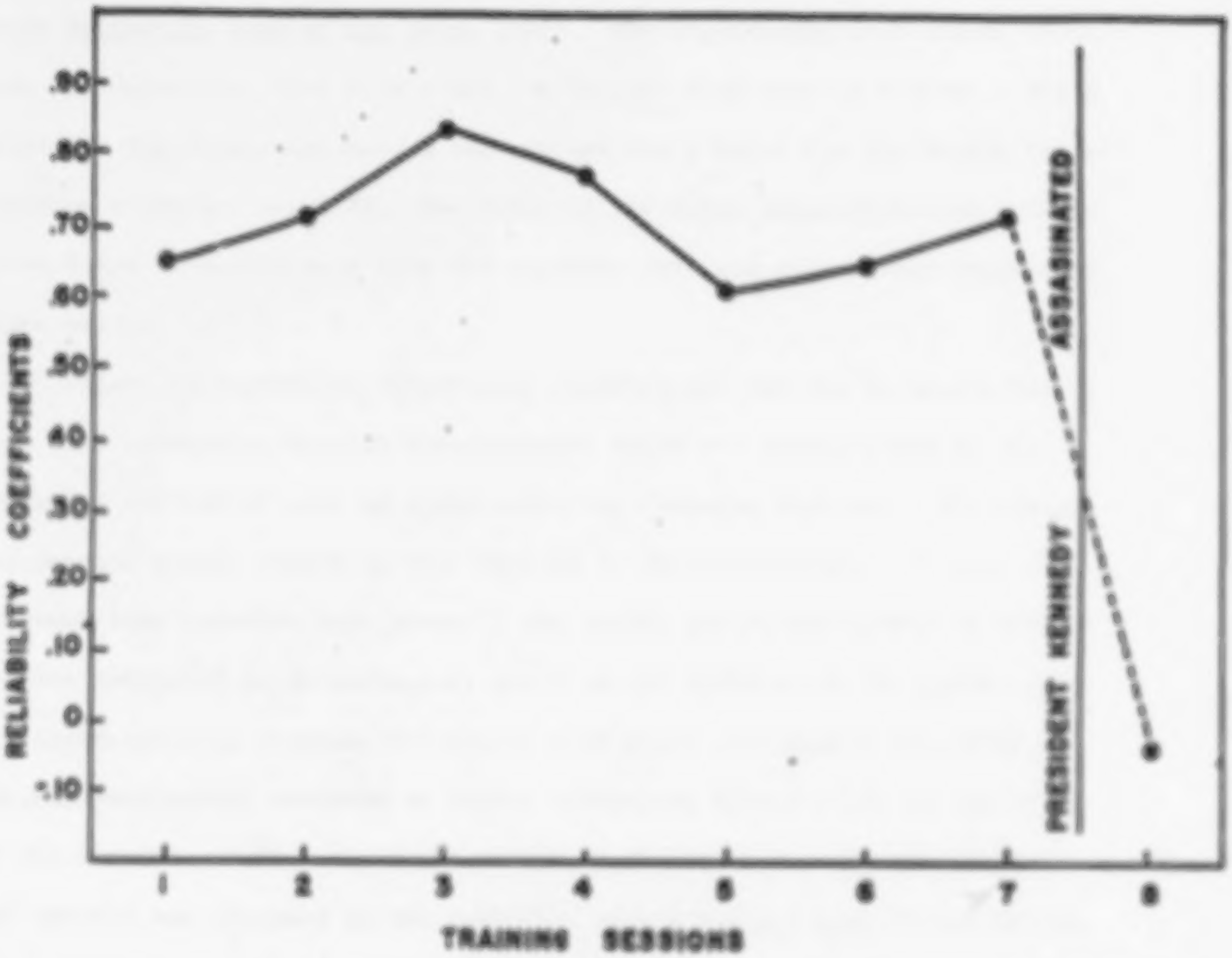


Fig. 14. Reliability coefficients for the structure dimension of the Leadership Opinion Questionnaire for training sessions before and after the assassination of President Kennedy (from Ayers, 1964).

test was 1.10, the *F* values for the Walker-Lev tests 1 and 3 were .03 and 2.69, respectively, and the *t* value for the Double Extrapolation Technique was .82. The value of the first autocorrelation coefficient based on departures from the total series regression line was .15.

Figure 13 represents the number of hospitalized mental patients in the United States before and after the advent of the use of tranquilizing drugs (from *Britannica Book of the Year, 1965*). The significant statistical tests were the Walker-Lev Test 2 (F = 213.75) and the Mood test (t = 7.40). The *F* value for the Walker-Lev test 3 was .26 and the *t* value for the Double Extrapolation technique was 1.38. The value of the first autocorrelation coefficient based on differences from the separate pre- and post-change regression lines was .52.

Figure 14 represents reliability coefficients for the Structure Dimension of a Leadership Opinion Questionnaire which was administered at the beginning and end of each of eight week-long training sessions. The training session groups ranged in size from 22 to 55 participants. Classes for the week-long sessions were given in the Spring and Autumn months of 1963 and are indicated in chronological order on the abscissa of the graph. For the eighth training session, pretesting took place on November 18, 1963, whereas posttesting occurred on Friday afternoon, November 22, at the close of the session. During the lunch period, 1:30-2:30 P.M., the participants had watched and listened to the memorable events being reported from Dallas, Texas (from Ayers, 1964). Since only one post-change point is given, only the Mood test of significance is applicable. The obtained *t* value for the Mood test was 6.24 with 5 *df*, thus, confirming the impression of effect. The value of the first autocorrelation coefficient based on differences from the separate pre- and post-change regression lines was .12.

As the above examples of actual time series illustrate, the tests of significance are not equally applicable to all time series data. The several possible time series of Figure 15 indicate this. The time series of A1 and A2 are instances of sustained post-change effect involving a change in intercept; whereas, B1 and B2 show an initial jump and subsequent return to pre-change conditions. If theoretical expectations are appropriate to using a series of points in the post-change period (as they would be in A1 and A2) the Walker-Lev Test 3 and Double Extrapolation Technique would be most suitable. These two tests would not be used in instances of theoretical expectations similar to B1 and B2 where only the Mood Test of the significance of the first-post change observation from a trend extrapolated from the pre-change observations would be appropriate.

Insert Figure 15 about here

A times series analysis seems most suitable in instances discussed above. In C1, C2, and C3 where the post-treatment effect involves a change in slope and, generally, of intercept, the possibility of a curvilinear relationship or cyclical trend is more difficult to rule out. A large number of pre- and post-treatment observations would be helpful in eliminating these possibilities. All three statistical tests and, in particular, the Walker-Lev Test 1, may be used, but they may yield conflicting results. For example, the double extrapolation technique can be used to indicate if the pre- and post-regression lines coincide at time t_o midway between the last pre-change and first post-change point. However, extension of the regression lines of C2 will show that the two lines coincide at time t_o, although pre- and post-regression lines are dissimilar.

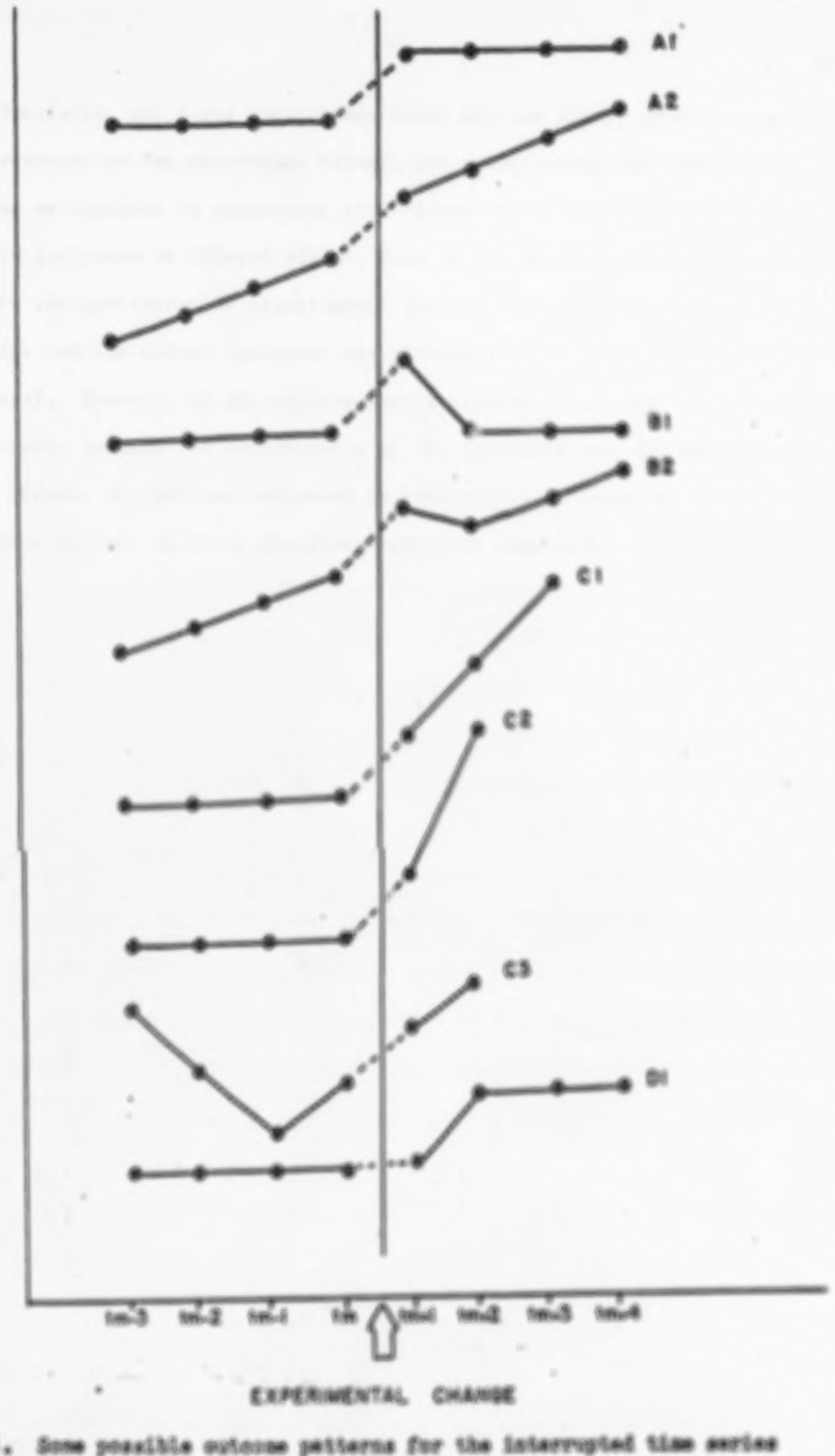


Fig. 35. Some possible outcome patterns for the interrupted time series experiment.

Similarly, the first post-change point may not differ greatly from that predicted by the pre-change values, yet a continuing and substantial increase or decrease in subsequent post-change values may suggest an effect.

In instances of delayed effect, such as D1, ambiguity is introduced into the interpretation of significance (as the time interval between the treatment and its effect increases the plausibility of rival hypotheses also increases). However, if the experimenter specifies in advance the exact relationship between the introduction of the treatment and the manifestation of its effect, the pattern indicated by time-series D1 could be almost as definitive as that in which immediate effect is expected.

References

Ayers, Arthur (Westinghouse Electric Corporation). Letter in *American Psychol.*, *19*, 1964, p. 353.

Britannica Book of the Year, 1965. Chicago, Encyclopedia Brittanica, 1965, p. 626.

Campbell, D. T. From description to experimentation. In C. W. Harris (Ed.), *Problems in measuring change*. Madison, Wis.: University of Wisconsin Press, 1963, pp. 212-242.

Campbell, D. T., and Stanley, J. S. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963, pp. 171-246.

Holtzman, W. H. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change*. Madison, Wis.: University of Wisconsin Press, 1963, pp. 199-211.

Mood, A. F. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1950.

Sween, Joyce A., and Campbell, D. T. The interrupted time series as quasi-experiment: three tests of significance. Northwestern University, 1966.

Uniform Crime Reports for the United States. Issued by the Federal Bureau of Investigation, United States Dept. of Justice, Washington, D.C.

Walker, Helen M., and Lev, J. *Statistical inference*. New York: Holt, 1953.

Footnotes

[1] A preliminary investigation of the Clayton Test using 100 sets of generated time series (total N = 40, true line slope - 1.00, total error variance = 5.00) yielded alpha values greater than the expected one and five percents when no dependency between points was built in (independent error). The percents of significant instances exceeding the five percent tabled critical value of F were 17%, 36%, 53%, and 61% for independent error, lag 1, lag 2, and lag 3 correlated errors, respectively. In lieu of this preliminary finding, the Clayton test was not considered a useful significance test in the interrupted time series situation and it was eliminated from further Monte Carlo investigation.

In a similar preliminary investigation, the Walker-Lev Test 1 yielded 4% (independent error), 16% (lag 1 error), 26% (lag 2 error), and 34% (lag 3 error) false-positive instances above the five percent tabled critical value of F. Although the Walker-Lev Test 1 yielded results similar to those obtained in a preliminary investigation of the Walker-Lev Test 3, it was eliminated in the final 1000 set investigation. Preference was given to the Walker-Lev Test 3 because of the more restricted usefulness of Test 1. Since the Walker-Lev Test 1 is a test of slope differences only it is more difficult to rule out rival curvilinear hypothesis in cases of significance.

The Walker-Lev Test 2 was not used because a test of the null hypothesis of zero slope is not of general interest as a test of significance in the time series situation.

THE INTERRUPTED TIME SERIES AS QUASI-EXPERIMENT: THREE TESTS OF SIGNIFICANCE*

A Fortran Program for the CDC 3400 Computer

Joyce Sween and Donald T. Campbell

August, 1965

Distributed by the Vogelback Computing Center
Northwestern University as NUC 0049 TIMEX

---

*Supported by Project C998, Contract 3-20-001, Educational Media Branch Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

Table of Contents

Page

Interrupted Time Series Design 2

Three Statistical Tests of Significance. 2

    Mood Test 2

    Walker-Lev Tests. 4

    Double Extrapolation Technique. 6

Considerations for Using Tests of Significance 7

Use of Tests When Assumption of Uncorrelated Error is Not Met. . . . 11

Autocorrelation Coefficients 18

General Program Description. 19

Program Limitations. 21

Running Time Estimation. 21

Card Preparation 22

Card Order 23

Fortran Program and Sample Output 24

References 36

Appendix I: Interpretation of Data Plot with Missing
Regression Points. 37

Appendix II:Program Modifications for Fortran II and IV. 39

The Interrupted Time Series as Quasi-Experiment: Three Tests of Significance

A Fortran Program for the CDC 3400 Computer

Joyce Sween and Donald T. Campbell

Studies are classified as quasi-experimental where random assignment to treatments is not possible. Campbell and Stanley (1963) have assembled a considerable number of such designs, of which the interrupted time series is one. A discussion of the time series design as a quasi-experimental method and the problems of statistical tests and of internal validity for this design are presented in Campbell and Stanley (1963) and Campbell (1963).

The essence of the interrupted time series experiment is the presence of a periodic measurement process in some individual or group and the introduction of an experimental change into this time series of measurements. The kind of time series that is considered here is a finite sequence of observations which are real-valued measurements obtained at n successive equally spaced points in time. The experimental change or treatment is assumed to occur between two consecutive measurements. Measurements obtained prior to the treatment are referred to as pre-change values; measurements obtained after the treatment are post-change values. It can be diagrammed as follows:

Interrupted Time Series Design

Time:	t_1	$t_2 \cdots t_m$		t_{m+1}	$t_{m+2} \cdots t_n$
Observations at time t_i:	y_1	$y_2 \cdots y_m$	↑	y_{m+1}	$y_{m+2} \cdots y_n$
	pre-change values			post-change values	
			experimental change		

Three Statistical Tests of Significance

If the treatment has produced an effect, a discontinuity of measurement is recorded in the time series. It is desirable, of course, to have available some statistical test of significance that would distinguish such a treatment effect from purely random fluctuation. A recent Monte Carlo investigation (Sween and Campbell, 1965) indicated that the following three statistical tests were useful in the interrupted time series design:

(a) A *t* test (Mood, 1950, pp. 297-298) for the significance of the first post-change observation from a value predicted by a linear fit of the pre-change observations.

(b) An *F* test (Walker and Lev, 1953, pp. 390-395) of the hypothesis that one regression line fits both pre- and post-change groups.

(c) A *t* test of the difference between pre-change and post-change regression estimates of the occasion value lying midway between the last pre-change point and the first post-change point (Mood, 1950, pp. 350-358; Walker and Lev, 1953, pp. 390-400).

Computational formulae for these three tests are described in detail on the following pages.

Mood Test. Mood (1950, pp. 297-298) provides a *t* test for the significance of the difference between the first post-change observation and a predicted value based upon a linear fit of the pre-change observations. The formula for *t* is given below:

$$t = \frac{(Y_0 - \hat{Y}_0)/\sigma_m}{\sqrt{m\hat{\sigma}^2/(m-2)\sigma^2}}$$

The computer program first obtains the pre-change regression estimates.

$$\hat{\beta} = \frac{\sum_{i=1}^{m} t_i Y_i - \left[\left(\sum_{i=1}^{m} t_i\right)\left(\sum_{i=1}^{m} Y_i\right)/m\right]}{\sum_{i=1}^{m} t_i^2 - \left(\sum_{i=1}^{m} t_i\right)^2/m}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{t}$$

where m = number pre-change observations

$i = 1, 2, \ldots m$

$\bar{y} = \sum_{i=1}^{m} Y_i/m$

$\bar{t} = \sum_{i=1}^{m} t_i/m$

An estimate, $\hat{Y}_0$, of the first post-change value is obtained.

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta}t_0$$

The difference between the estimated value and the obtained first post-change value, $Y_0 - \hat{Y}_0$, is used in the *t* test. However, $\sigma_m = \sqrt{\sigma^2\left[\frac{m+1}{m} + \frac{(t_0 - \bar{t})^2}{\sum_{i=1}^{m}(t_i - \bar{t})^2}\right]}$ and an estimate, $\hat{\sigma}^2$, of σ^2 is given by $\hat{\sigma}^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{\alpha} - \hat{\beta}t_i)^2$. Thus, the computation formula for the *t* test of the difference between the obtained and predicted points is given by

$$t = \frac{Y_0 - \hat{Y}_0}{\sqrt{\left[\frac{m+1}{m} + \frac{(t_0 - \bar{t})^2}{\sum(t_i - \bar{t})^2}\right]\left[\sum_{i=1}^{m}\frac{(Y_i - \hat{\alpha} - \hat{\beta}t_i)^2}{m-2}\right]}}$$

The denominator is the standard error of the difference from *t*. The *df* for *t* is (m-2).

Walker-Lev Tests.--Walker and Lev (1953, pp. 390-395) provided the following tests of significance. These tests are useful in determining whether differences exist between the regression equations for pre- and post change groups.

Test One.

This is a test of the hypothesis of common slope. The F ratio is given by $F = \frac{S_1}{S_2} \cdot \frac{(N-4)}{1}$

where S_1 = sum of squares of the common within groups regression estimates from the separate group regression estimates

$$\sum_{i=1}^{2} \sum_{j=1}^{N_i} \left[(\hat{\alpha}_i + \hat{\beta}_i t_{ij}) - (\hat{\alpha}_{w_i} + \hat{\beta}_w t_{ij}) \right]^2$$

S_2 = sum of squares of the obtained occasion values from the separate group regression estimates

$$\sum_{i=1}^{2} \sum_{j=1}^{N_i} \left[(Y_{ij} - (\hat{\alpha}_i + \hat{\beta}_i t_{ij}) \right]^2$$

N = total number of occasions in time series

i=1,2 [1=prechange group; 2=postchange group]

j = 1, 2. . . N_i

Formulas for the common within groups slope and intercept are

$$\hat{\beta}_w = \frac{CTY_w}{CTT_w}$$

$$\hat{\alpha}_{w_i} = \bar{Y}_i - \hat{\beta}_w \bar{t}_i$$

Formulas for the separate group slopes and intercepts are

$$\hat{\beta}_i = \sum_{j=1}^{N_i} t_{ij} Y_{ij} - \left(\sum_{j=1}^{N_i} t_{ij}\right)\left(\sum_{j=1}^{N_i} Y_{ij}\right) / N_i$$

$$\hat{\alpha}_i = \bar{Y}_i - \hat{\beta}_i \bar{t}_i$$

The computational formulae for the numerator and denominator sum of squares for the F ratio are

$$S_1 = \sum_{i=1}^{2} \frac{CTY_i}{CTT_i} - \frac{C^2 TY_w}{CTT_w}$$

$$S_2 = CYY_w - \sum_{i=1}^{2} \frac{C^2 TY_i}{CTT_i}$$

where

$$CTT_i = \sum_{j=1}^{N_i} t_{ij}^2 - \left[\left(\sum_{j=1}^{N_i} t_{ij}\right)^2 / N_i\right]$$

$$CTY_i = \sum_{j=1}^{N_i} t_{ij} Y_{ij} - \left[\sum_{j=1}^{N_i} t_{ij} Y_{ij} / N_i\right]$$

$$CYY_i = \sum_{j=1}^{N_i} Y_{ij}^2 - \left[\left(\sum_{j=1}^{N_i} Y_{ij}\right)^2 / N_i\right]$$

$$CTT_w = \sum_{i=1}^{2} CTT_i$$

$$CTY_w = \sum_{i=1}^{2} CTY_i$$

$$CYY_w = \sum_{i=1}^{2} CYY_i$$

The resulting *F* ratio is evaluated with 1 and N-4 degrees of freedom.

Test Two:

This is a test of the hypothesis that the slopes of the pre- and post-change groups are equal to zero when it has been established that $\hat{B}_i = \hat{B}_w$ (that is, *F* of Test One is not significant). The variance ratio is given by

$$F = \frac{C^2TY_w}{CTT_w \cdot CYY_w - C^2TY_w} \cdot \frac{(N-4)}{1}$$

and is evaluated with 1 and N-4 degrees of freedom.

Test Three:

This is a test of the hypothesis that a single regression line fits both the pre- and post-change groups. The *F* ratio is given by $F = \frac{S_b}{S_w} \cdot \frac{(N-3)}{1}$

where S_b = sum of squares of common within groups regression estimates from the total series regression estimates

$$\sum_{i=1}^{2} \sum_{j=1}^{N_i} \left[(\hat{\alpha}_{w_i} + \hat{B}_w t_{ij}) - (\hat{\alpha}_T + \hat{B}_T t_{ij})\right]^2$$

S_w = sum of squares of obtained occasion values from the common within groups regression estimates

$$\sum_{i=1}^{2} \sum_{j=1}^{N_i} \left[Y_{ij} - (\hat{\alpha}_{w_i} + \hat{B}_w t_{ij})\right]^2 .$$

Formulas for the total series slope and intercept are

$$\hat{\beta}_T = \frac{CTY_T}{CTT_T}$$
$$\hat{\alpha}_T = \bar{y} - \hat{\beta}_T \bar{t}$$

The computational formulae for the numerator and denominator sum of squares for the *F* ratio are

$$S_b = CYY_B + \frac{C^2TY_w}{CTT_w} - \frac{C^2TY_T}{CTT_T}$$
$$S_w = S_1 + S_2$$

where

$$CTY_T = \sum_{i=1}^{2} \sum_{j=1}^{N_i} t_{ij} y_{ij} - \left[\left(\sum_{i=1}^{2} \sum_{j=1}^{N_i} t_{ij} \right) \left(\sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{ij} \right) / N_i \right]$$
$$CYY_T = \sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{ij}^2 - \left[\left(\sum_{i=1}^{2} \sum_{j=1}^{N_i} y_{ij} \right)^2 / N \right]$$
$$CTT_T = \sum_{i=1}^{2} \sum_{j=1}^{N_i} t_{ij}^2 - \left[\left(\sum_{i=1}^{2} \sum_{j=1}^{N_i} t_{ij} \right)^2 / N \right]$$
$$CYY_B = CYY_T - CYY_w$$

The resulting *F* ratio is evaluated with 1 and N-3 degrees of freedom.

"Double Extrapolation" Technique.--This test (Mood, 1950, pp. 350-358; Walker and Lev, 1953, pp. 399-400) is concerned with the significance of the difference between two separate regression estimates of the y-value for time t_0 which lies midway between the last pre-change point and the first post-change point. Reference to the interrupted Time Series Design diagrammed on page 2 will indicate that this point lies midway between t_m and t_{m+1}. Assuming that the points are equally spaced in time, t_0 is equal to $t_{m+.5}$.

The first regression estimate for y_0 is obtained from the pre-change values; the second regression estimate for y_0, from the post-change values.

The formulas for the two regression estimates are

$$\hat{\beta}_i = \frac{\sum_{j=1}^{N_i} t_{ij} y_{ij} - \left[\left(\sum_{j=1}^{N_i} t_{ij}\right)\left(\sum_{j=1}^{N_i} y_{ij}\right)/N_i\right]}{\sum_{j=1}^{N_i} t_{ij}^2 - \left(\sum_{j=1}^{N_i} t_{ij}\right)^2/N_i}$$

$$\hat{\alpha}_i = \bar{y}_{ij} - \hat{\beta}\bar{t}_{ij}$$

where $i = 1, 2$ [1 = prechange group; 2 = post-change group]
N_i = number occasions in group i
$j = 1, 2, \ldots N_i$

Thus, the two estimates for y_o are given by $\hat{y}_{o_i} = \hat{\alpha}_i + \hat{\beta}_i t_o$. The difference between these two estimates can be evaluated by the _t_ ratio $t = \frac{|\hat{y}_{o_1} - \hat{y}_{o_2}|}{S_D}$ and its significance determined by reference to a _t_ table with $N_1 + N_2 - 4$ degrees of freedom. The standard error of the difference for _t_ depends upon the variability of the pre- and post-change values, N_1 and N_2, the relation between t and y, and the distance of t_o from $\bar{t}_1$ and $\bar{t}_2$. The formula for the standard error is given in Walker and Lev as

$$S_D = \sqrt{\frac{\left[\sum_{i=1}^{2}\left(CYY_i - \frac{C^2TY_i}{CTT_i}\right)\right]\left[\frac{N_1+N_2}{N_1N_2} \quad \sum_{i=1}^{2}\frac{(t_o - \bar{t}_i)^2}{CTT_i}\right]}{N_1+N_2-4}}$$

where

$$CYY_i = \sum_{j=1}^{N_i} y_{ij}^2 - \left(\sum_{j=1}^{N_i} y_{ij}\right)^2/N_i$$

$$CTT_i = \sum_{j=1}^{N_i} t_{ij}^2 - \left(\sum_{j=1}^{N_i} t_{ij}\right)^2/N_i$$

$$CTY_i = \sum_{j=1}^{N_i} t_{ij} y_{ij} - \left(\sum_{j=1}^{N_i} t_{ij} \sum_{j=1}^{N_i} y_{ij}\right)/N_i$$

Considerations for Using Tests of Significance

The following points should be considered in utilization of these three tests of significance.

(1) The three tests are linear tests of intercept and/or slope. In instances of significance, curvilinear relationships must be ruled out as rival hypotheses explaining the post-change effect. Thus, a plot of the time series should be examined to determine by visual inspection if other than a linear fit seems plausible. (The present FORTRAN program provides a data plot of the time series).

(2) Both pre-treatment and post-treatment measurements are necessary for a time series analysis. In general, there should be sufficient pre-change and post-change data available to establish a trend. For the Mood test only one post-change point is required, but additional post-change data would be helpful in ruling out rival hypotheses.

(3) The three tests of significance are not equally applicable to all time series analyses, as the following examples should illustrate. The time series of A1 and A2 of Figure 1 are instances of sustained post-change effect involving a change in intercept whereas B1 and B2 show an initial rise and subsequent return to pre-change conditions. If theoretical expectations are appropriate to using a series of points in the post-change period (as they would be in A1 and A2) the Walker-Lev Test 3 and "double extrapolation" technique would be most suitable. These two statistics would not be used in instances of theoretical expectations similar to B1 and B2 where only the Mood test of the significance of the first post-change observation from a trend extrapolated from the pre-change observations would be appropriate.

Insert Figure 1 about here

A time series analysis appears to be most applicable in the instances cited above. In C1 and C2 where the post-treatment effect involves a change

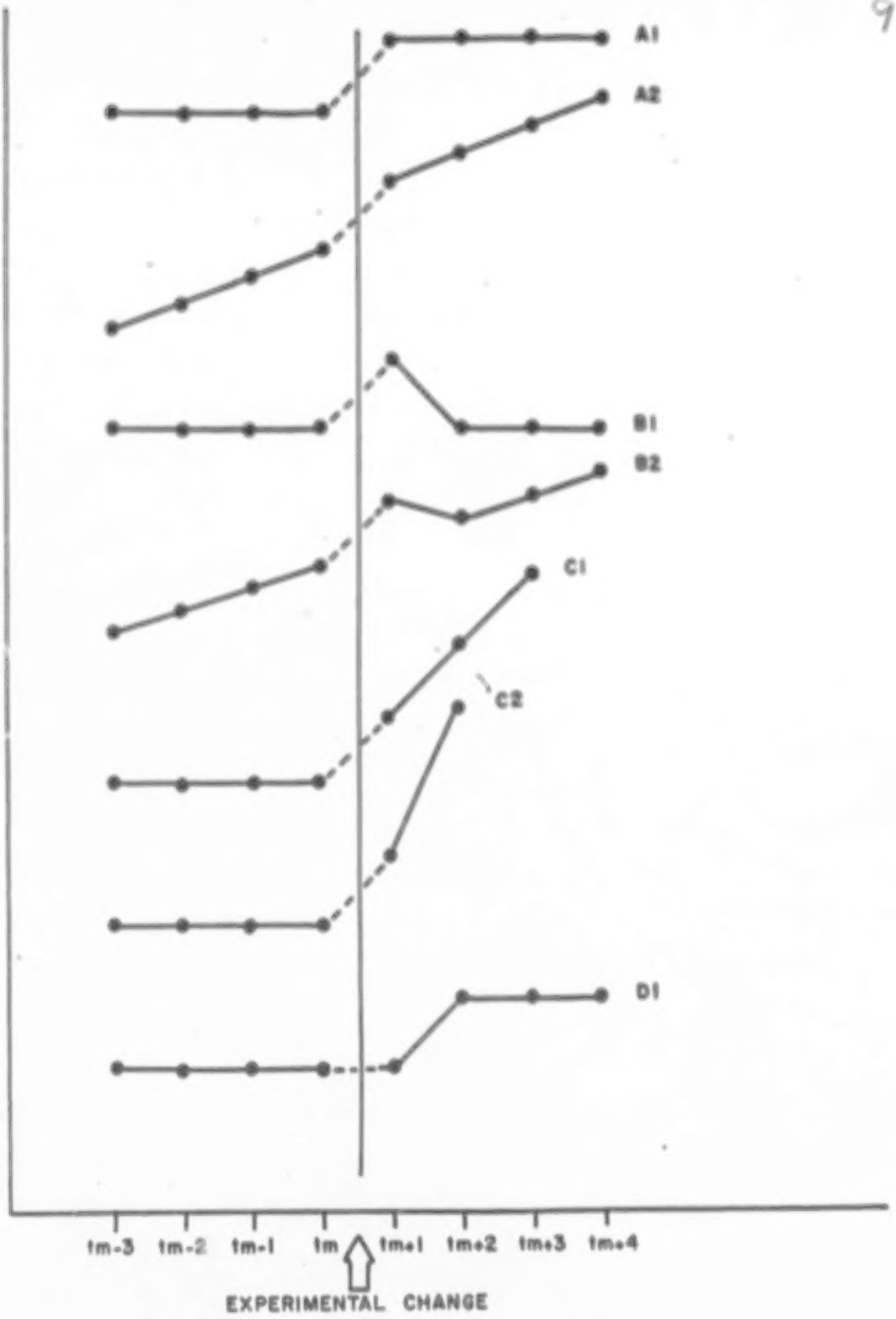


Fig. 1. Some possible outcome patterns for the interrupted time series experiment.

in slope and, generally, of intercept, the possibility of a curvilinear relationship is more difficult to rule out. Although all three statistical tests may be appropriate here (depending upon the theoretical expectations of the investigator), they may yield conflicting results. For example, the "double extrapolation" technique could be used to indicate if the pre- and post-regression lines coincide at time, t_0, midway between the last pre-change and first post-change point. However, in instances like C2 both regression lines coincide at time t_0 even though their slopes and intercepts are dissimilar.

In instances of delayed effect, such as D1, ambiguity is introduced into the interpretation of significance. For as the time interval between the treatment and its effect increases, the plausibility of rival hypotheses also increases. In using the tests of significance, it is essential that the investigator specify in advance the time relationship between the introduction of the treatment and the manifestation of its effect. If a delayed effect is hypothesized, all points prior to the effect should be regarded as pre-change points. (The Occasion Card used as input to the program must also be punched accordingly.)

(4) Statistical models on which these tests of significance are based make the minimal assumptions that the errors have equal variances and expected values of zero. In addition, the errors are assumed to be uncorrelated. However, the observed values may not always exhibit independent errors. It may frequently occur that measurements made closely together in time are more strongly related. When such dependence between the measurements exists, the assumption of independence is not met. Although this circumstance does not prevent utilization of the tests of significance with interrupted time

series data, the Monte Carlo study (Sween and Campbell, 1965) indicated that if correlated error exists, use of these tests with the tabled critical values will produce significant results by chance alone *more than* one percent (or five percent) of the time. The degree to which the one percent and five percent levels of significance were exceeded depended upon the amount of autocorrelation between points. This was found to hold particularly for the Walker-Lev Tests and the "double extrapolation" technique. In lieu of these findings, the corrections described below are advised in instances where the assumption of independent error is not justified.

Use of Tests When Assumption of Uncorrelated Error is not Met

On the bases of the Monte Carlo study, it is suggested that when a significant autocorrelation exists, the critical values of F (Walker-Lev tests) and of t ("double extrapolation" technique and Mood test) be increased to approximate more closely the 1% and 5% levels of significance. Only a slight correction, if any, is needed for the Mood test. Reference to the graphs in Figures 2-7 should facilitate the approximations. The new critical values can be determined by interpolating on the abscissa using the first autocorrelation coefficient (i.e., for a lag of one step) and then interpolating for the total N or pre-change N (Mood Test) of the series on the face of the graph. The new critical value is read on the ordinate.

Insert Figures 2-7 about here

As the graphs indicate, when observations are not correlated the critical values yielding significant results one and five percent of the time by

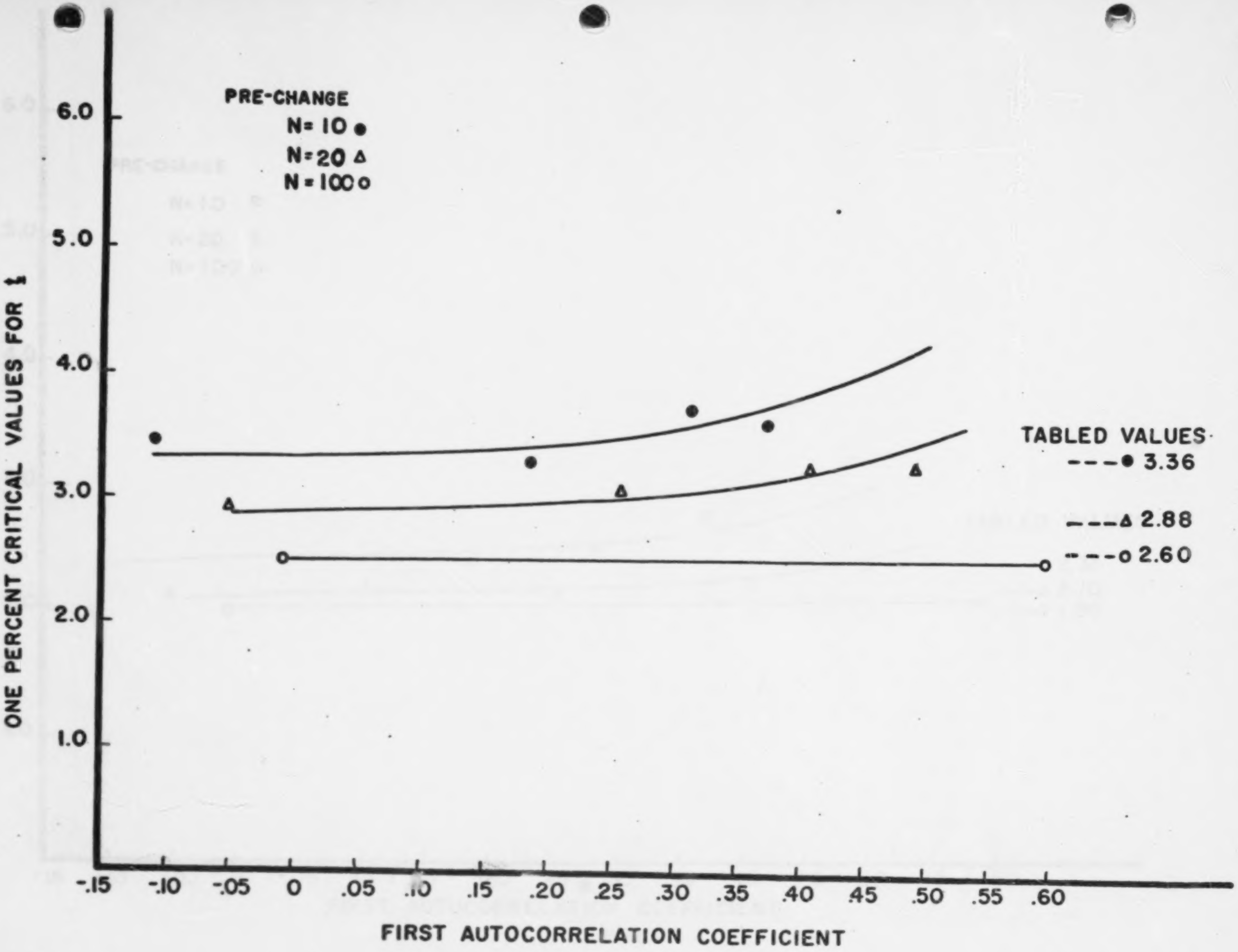


Fig. 2. Mood test: Obtained one percent critical values of t as a function of the first autocorrelation coefficient.

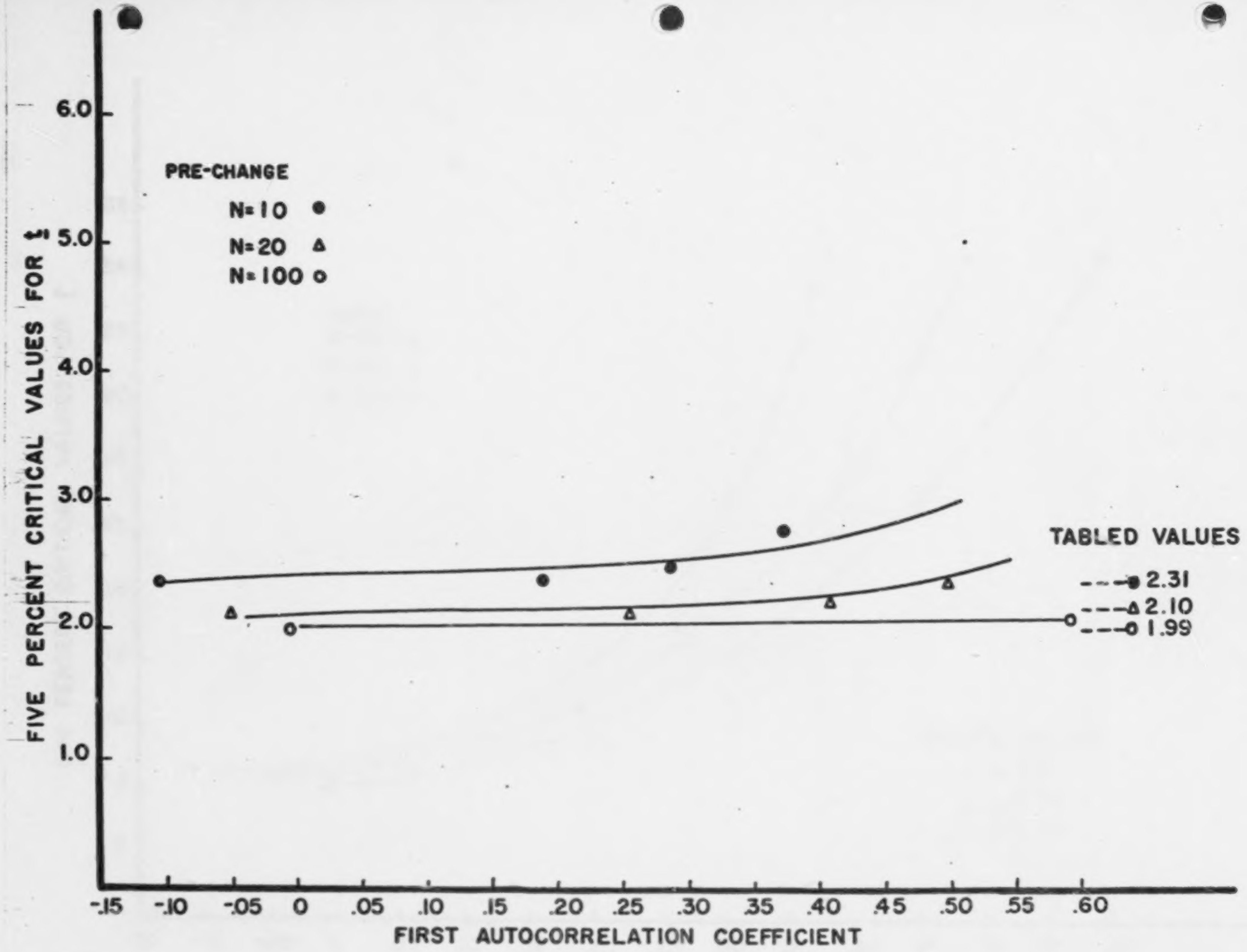


Fig. 3. Mood test: Obtained five percent critical values of t as a function of the first autocorrelation coefficient.

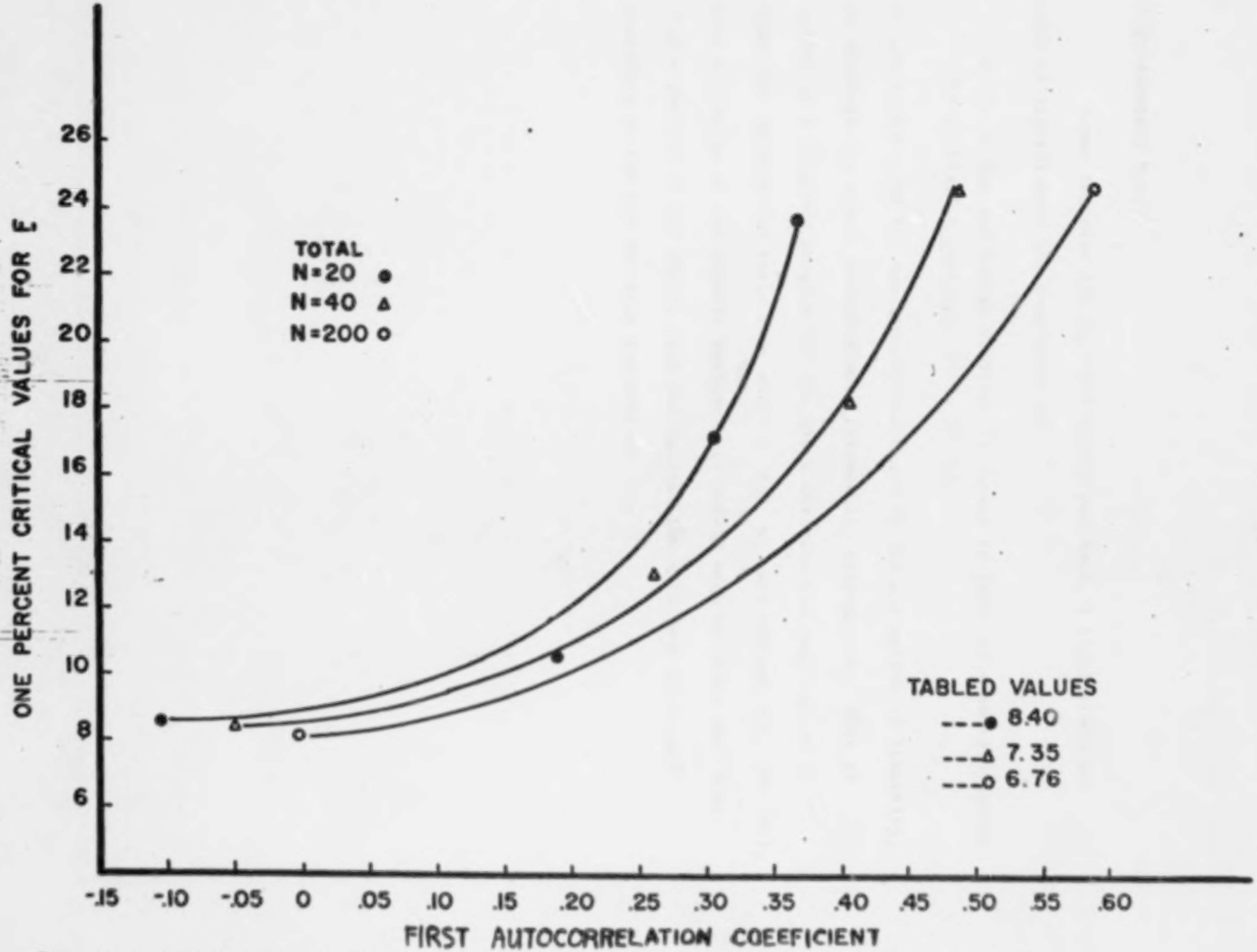


Fig. 4. Walker-Lev Test 3: Obtained one percent critical values of F as a function of the first autocorrelation coefficient.

Supplementary Note:

Since the research reported herein was done, a highly relevant test of significance has been reported:

> G. E. P. Box and George C. Tiao, "A change in level of a non-stationary time series," Biometrika, 1965, 52, 181-192.

In particular, the Box and Tiao approach avoids the assumption of linearity, in exchange for other, probably more reasonable, assumptions. What is needed is a computer program for the exact distribution computation by numerical methods for their test when δ and γ_o are unknown (pp. 189-191), plus a testing of the formula against null models such as those used here, plus a testing of our linear formulas against the null data generated according to the Box and Tiao assumptions.

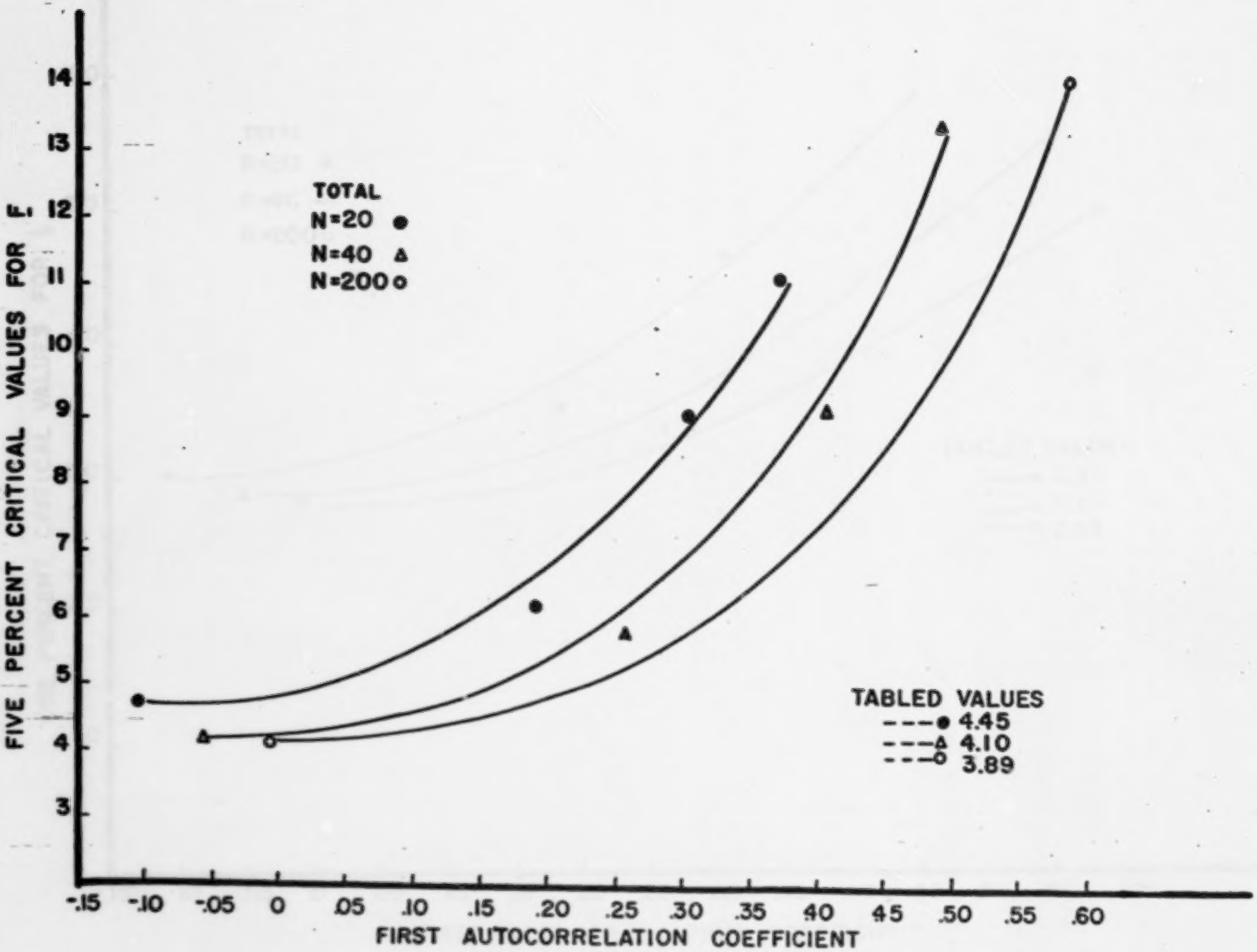


Fig. 5. Walker-Lev Test 3: Obtained five percent critical values of *F* as a function of the first autocorrelation coefficient.

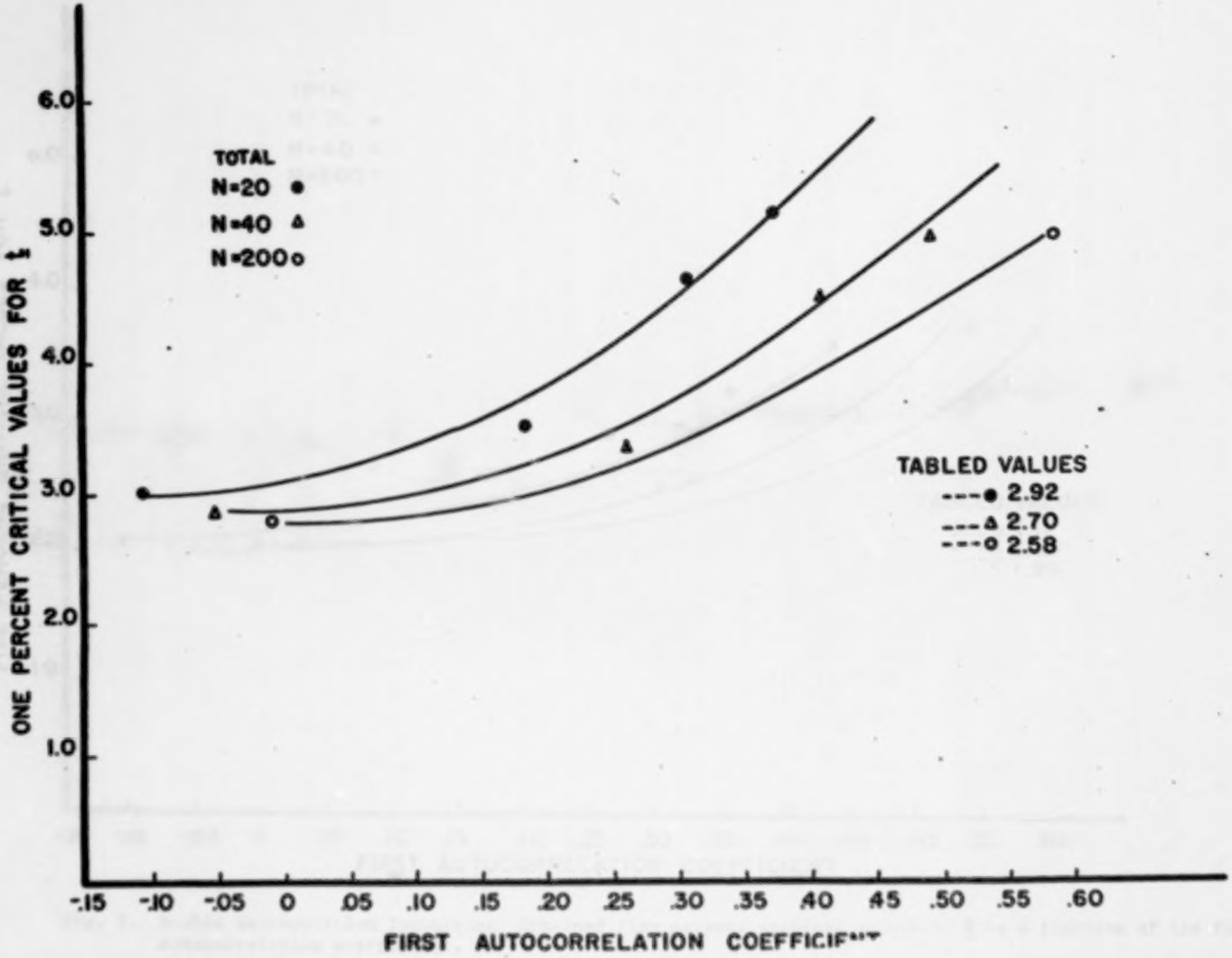


Fig. 6. Double Extrapolation Technique: Obtained one percent critical values of *t* as a function of the first autocorrelation coefficient.

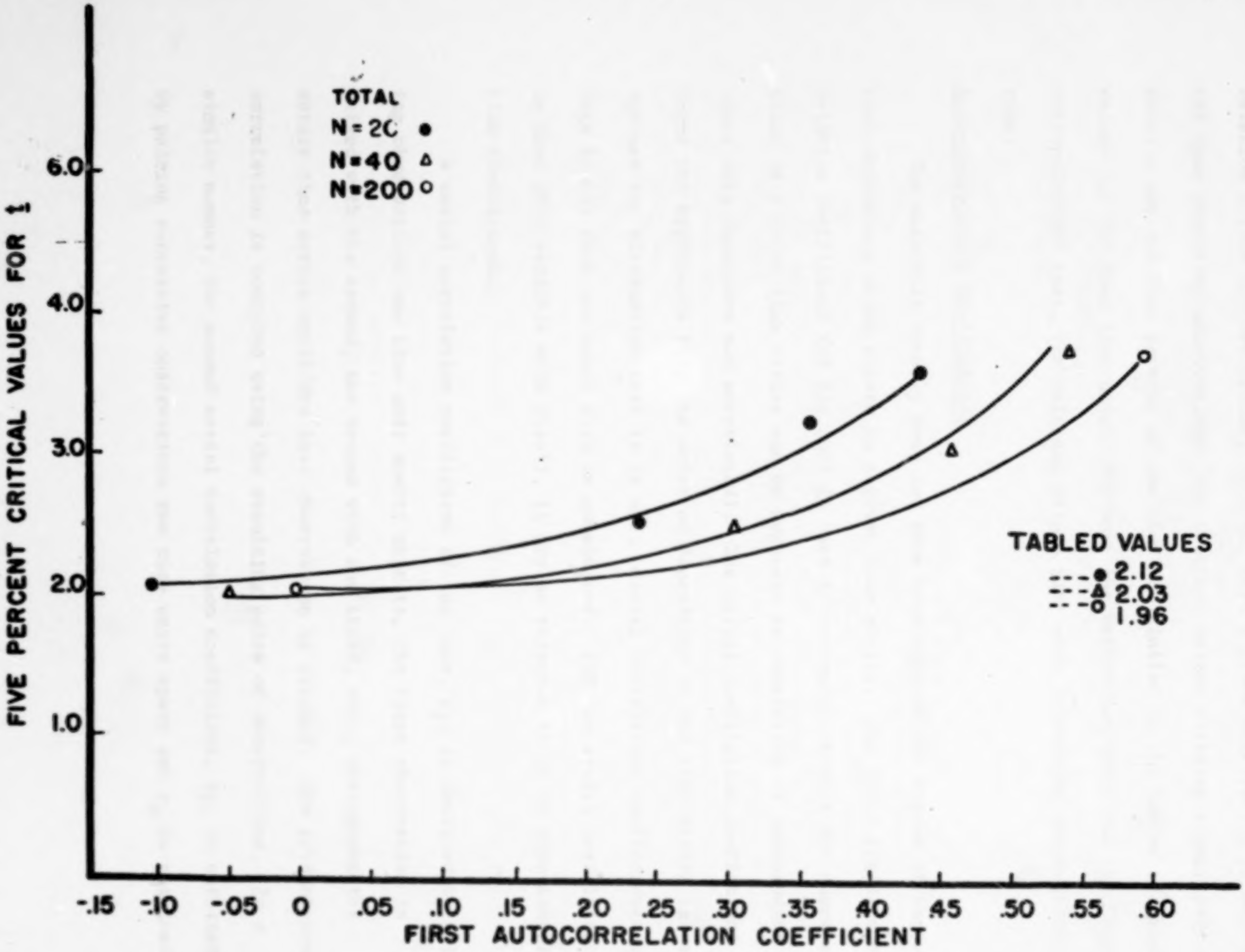


Fig. 7. Double Extrapolation Technique: Obtained five percent critical values of *t* as a function of the first autocorrelation coefficient.

chance alone are, in general, similar to the tabled critical values for all three tests of significance. However, when a significant proximal autocorrelation exists between points, that is, when a given observation is dependent upon preceding observations, the critical values yielding significant results one and five percent of the time are similar to the tabled critical values for the Mood test only. For both the Walker-Lev tests and the "double extrapolation" test, the critical values rise with increasing autocorrelation.

Auto-correlation Coefficients

The scientist usually does not have knowledge of the degree of sequential dependency which exists in a given time series. The first serial correlation coefficient (of lag one) is used to determine whether the observations of a given time series can be regarded as consisting of independent error only (measures not correlated). The serial correlation coefficient tests the hypothesis that the order of dependence in the time series is zero against the alternative that it is one. Serial correlation coefficients of lags higher than one could also be considered. (If the serial correlation is that of a variable with itself, it may be referred to as an auto-correlation coefficient.)

A serial correlation coefficient of lag one, r_1, is obtained by pairing observations one time unit apart; that is, the first observation is paired with the second, the second with the third, etc., throughout the entire time series until the last observation is reached. The product-moment correlation is computed using the resulting pairs of observations. In a similar manner, the second serial correlation coefficient, r_2, is obtained by pairing successive observations two time units apart and r_m is obtained by pairing observations m units apart. The basic formula for r_a is given below:

$$r_a = \frac{n\sum_{i=1}^{n} y_i y_{i+a} - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} y_{i+a}\right)}{\sqrt{\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]\left[n\sum_{i=1}^{n} y_{i+a}^2 - \left(\sum_{i=1}^{n} y_{i+a}\right)^2\right]}}$$

where N = total number of occasions in the time series

n = N-a

i = 1,2,...n

When the slope of the regression line is not equal to zero, the presence of trend will increase the values of the autocorrelation coefficients. Thus, in this instance, interest would be in the autocorrelation of departures from the line of best fit for the total series. In the instance of a true effect (slopes and/or intercepts of pre- and post-change groups differ), the autocorrelation of the differences from the separate regression lines for the pre- and post-change groups will give the better estimate of the serial dependency between points.

Program Description

The computer program TIMEX is designed to perform the following statistical tests on data of the interrupted time series form:

1. Mood test
2. Walker-Lev covariance tests
3. "double extrapolation"

The tests desired are indicated on card input to the program (see card preparation).

The program also calculates autocorrelation coefficients for lags 1 through 4. The autocorrelations are also given for the departures from the regression line based on the total series and from the separate regression lines for the pre- and post-change values.

A data plot of the time series may also be obtained. In the data plot, the pre-change and post-change observation values are indicated by separate plotting characters. Initial and terminal points for regression lines based on the pre-change and post-change values and on the total series are indicated on the plot, if desired.

In addition to the _t_ or _F_ values of each statistical test of significance, the information outlined below may also be obtained. This option must be indicated on card input.

(a) Mood Test

*standard error for _t_ test
*estimated y-value for the first post-change point
*obtained y-value for the first post-change point
*slope and intercept for the pre-change points
*variance of the obtained pre-change points from the regression estimated values

(b) Walker-Lev Tests

Test 1
*numerator and denominator sum of squares for F
*regression estimates (slope and intercept) for pre-change values
*estimated y-value at experimental change (that is, midway between last pre-change and first post-change points) based upon pre-change regression estimates
*regression estimates for post-change values
*estimated y-value at experimental change based upon post-change regression estimates
*common within groups regression estimates for the pre-change values
*estimated y-value at experimental change based upon the above regression estimates
*common within groups regression estimates for post-change values
*estimated y-value at experimental change based upon the above regression estimates

Test 2

*numerator and denominator sum of squares for _F_

Test 3

*numerator and denominator sum of squares for F
*regression estimates (slope and intercept) for the total series
*estimated y-value at experimental change based upon the above regression estimates

(c) Double Extrapolation technique for Mood Test

*standard error for t test
*regression estimates for pre-change values
*variance of the obtained pre-change points from the values predicted on basis of the above regression estimates
*estimated y-value at experimental change based upon pre-change regression estimates
*regression estimates for the post-change values
*variance of the obtained post-change points from the values predicted on the basis of the above regression estimates
*estimated y-value at experimental change based upon post-change regression estimates

Program Limitations

The total number of occasion values must not be greater than 400. This is an arbitrary limit; the program could be modified to handle more than 400 points.

The data plot will not indicate more than 25 pre-change data points and/or 25 post-change data points.

Running Time Estimation

The following running times are reported to illustrate the relationship between size of sample and length of computer run.

Total N	Running Time*	Pages of Output
29	.09 min.	2
200	.10 min.	3

---

*Three Tests of Significance and Data Plot.

Card Preparation

The format and meaning of cards that are input to this program are described below.

a. Title Card

Col. 1-80	Identifying alphanumeric information. This information will appear as a heading on the print-out and on the data plot.

b. Occasion Card

Col. 1-3	Total number of occasions in the time series. The maximum number of occasions allowed is 400. The number of occasions in the pre- and post-change groups must be specified in Col. 4-9 below. The program assumes the y or occasion values occur at equalintervals in time.
Col. 4-6	Number of occasions in the pre-change group.
Col. 7-9	Number of occasions in the post-change group.

c. Test Specification Card

Col. 1	1 if calculations of serial correlation coefficients is desired. 0 (or blank) if you do not wish to have these calculations performed.
Col. 2	1 if Mood test is desired. 0 (or blank) if this statistical test is not desired.
Col. 3	1 if the Walker-Lev covariance tests are desired. 0 (or blank) if these tests are not desired.
Col. 4	1 if the "double extrapolation" technique for a double Mood test is desired. 0 (or blank) if this test is not desired.
Col. 5	1 if you wish to have the additional information specified on pages 13-14 printed out for each test of significance specified above. 0 (or blank) if you wish to have only the t or F values and the df for each test of significance. The additional information is not printed out.
Col. 6	1 if a data plot of input data points is desired. 2 if you wish terminal points for the regression lines based on pre-change and post-change values and on the total series to be indicated on the plot. 0 (or blank) if the data plot is not desired.

d. Variable Format Card

Col. 1-80 This card precedes the first data card. The format specification starts in the first card column and must be completed in 80 columns. The statement is enclosed by parentheses. The word FORMAT is not used, otherwise the format specification is as described in the FORTRAN manual. Only F and x conversion is allowed.

e. Data Cards

The data cards follow the Variable Format Card. The data arrangement on each card must be that given by the format specification on the Variable Format Card. However, it may sometimes occur that your last data card has less values than the preceding cards. This one exception is allowed.

Order of Cards

More than one problem can be processed consecutively. The following example illustrates the order of cards.

Title Card
Occasion Card
Test Specification Card
Variable Format Card
Data Cards

Title Card
Occasion Card
. . .
etc.

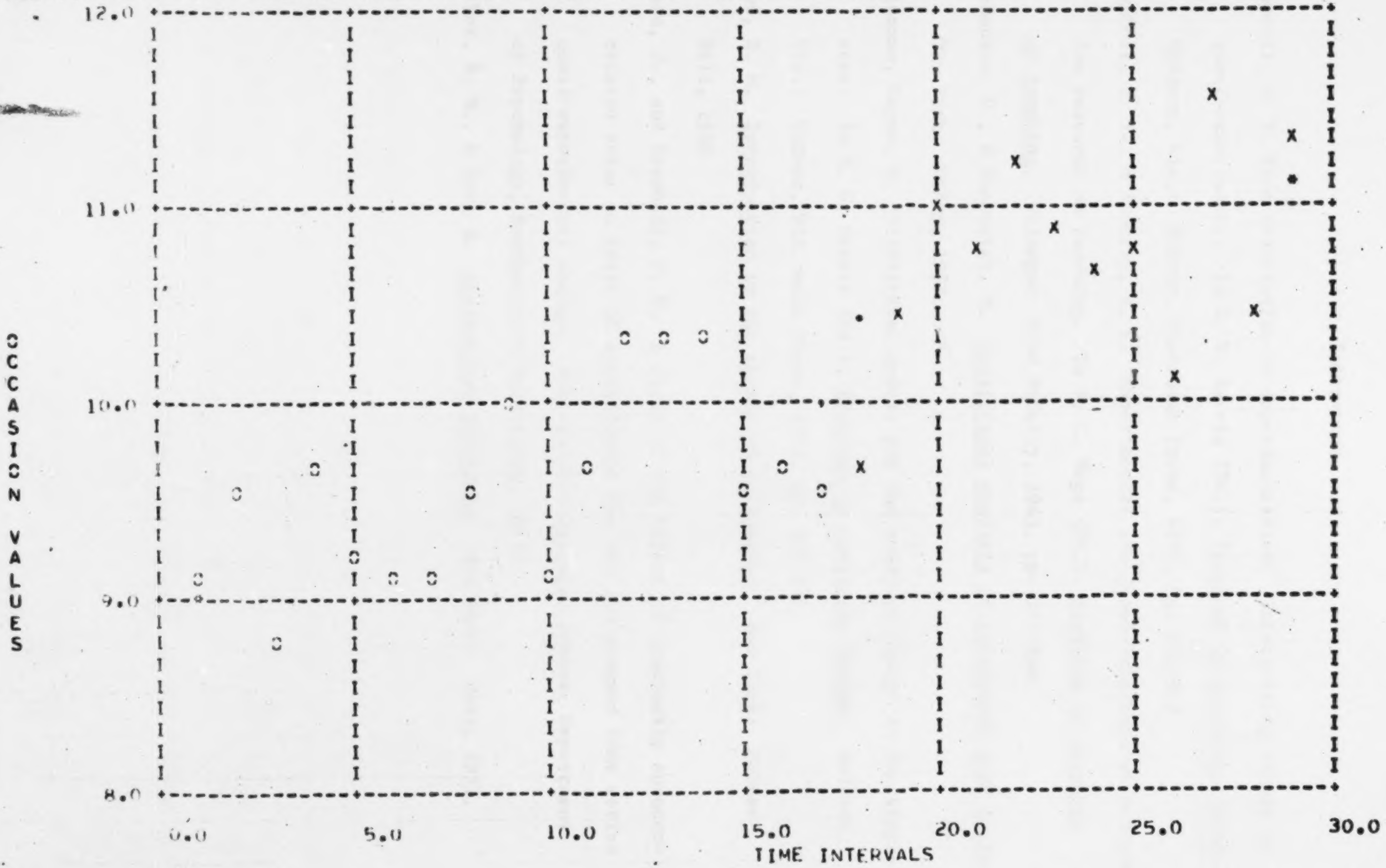
```
C        THE INTERRUPTED TIME SERIES AS QUASI-EXPERIMENT
C        THREE TESTS OF SIGNIFICANCE
C        PROGRAMMED BY JOYCE SWEEN,NORTHWESTERN UNIVERSITY
C
C        PLEASE SEE PROGRAM WRITEUP FOR SPECIFICATION CARDS REQUIRED
C        AS INPUT TO THIS PROGRAM AND FOR DESCRIPTION OF INPUT CARD ORDER.
C
      DIMENSION ITITLE(10),IFMDAT(10),T(400),Y(400),P(400),ZTOT(400),
     1ZPART(400)
      TYPE INTEGER BCD1,BCD2,BCDP1,BCDP2,BCDPT
C
C        READ SPECIFICATION AND DATA CARDS
C
  100 READ(60,10) ITITLE
      IF(EOF,60)999,110
  110 READ(60,11)NTOT,NPRE,NPOST
      IF(NTOT.GT.400) GO TO 91
      IF((NPRE+NPOST).NE.NTOT) GO TO 91
      READ(60,12)LAGCO,MOOD,LEV,MOOD2,MORE,NPLOT
      READ(60,10)IFMDAT
      READ(60,IFMDAT) (Y(I),I=1,NTOT)
      DO 402 I=1,NTOT
  402 T(I)=I
      NPRE1=NPRE+1 $PRE=NPRE $POST=NPOST $ TOT=NTOT
      IF((MOOD+LEV+MOOD2).EQ.0) GO TO 473
      WRITE(6,29)
      WRITE(6,30) ITITLE
      WRITE OUTPUT TAPE 6,130
      WRITE OUTPUT TAPE 6,120,(Y(I),I=1,NPRE)
      WRITE OUTPUT TAPE 6,131
      WRITE OUTPUT TAPE 6,120,(Y(I),I=NPRE1,NTOT)
      IF(LAGCO.NE.0)1100,409
C
C        CALCULATE AUTOCORRELATION COEFFICIENTS,IF OPTED
C
 1100 CALL LINE(1,NPRE,NPRE,T,Y,ALPHA1,BETA1)
      DO 1101 I=1,NPRE
 1101 P(I)=Y(I)-ALPHA1-(BETA1*T(I))
      CALL LINE(NPRE1,NTOT,NPOST,T,Y,ALPHA2,BETA2)
      DO 1102 I=NPRE1,NTOT
 1102 P(I)=Y(I)-ALPHA2-(BETA2*T(I))
      CALL SERIAL(NTOT,4,P,RSEP4)
      CALL SERIAL(NTOT,3,P,RSEP3)
      CALL SERIAL(NTOT,2,P,RSEP2)
      CALL SERIAL(NTOT,1,P,RSEP1)
      CALL LINE(1,NTOT,NTOT,T,Y,ALPHAT,BETAT)
      DO 1111 I=1,NTOT
 1111 P(I)=Y(I)-ALPHAT-(BETAT*T(I))
      CALL SERIAL(NTOT,4,P,RTOT4)
      CALL SERIAL(NTOT,3,P,RTOT3)
      CALL SERIAL(NTOT,2,P,RTOT2)
      CALL SERIAL(NTOT,1,P,RTOT1)
      CALL SERIAL(NTOT,4,Y,RSNR4)
      CALL SERIAL(NTOT,3,Y,RSNR3)
      CALL SERIAL(NTOT,2,Y,RSNR2)
      CALL SERIAL(NTOT,1,Y,RSNR1)
      WRITE(6,54)
      WRITE OUTPUT TAPE 6,46
      WRITE(6,55)RTOT1,RTOT2,RTOT3,RTOT4
      WRITE OUTPUT TAPE 6,47
      WRITE OUTPUT TAPE 6,55,RSEP1,RSEP2,RSEP3,RSEP4
      WRITE OUTPUT TAPE 6,45
      WRITE OUTPUT TAPE 6,55,RSNR1,RSNR2,RSNR3,RSNR4
  409 IF(MOOD.NE.0)700,418
C
C       MOOD TEST CALCULATIONS, IF OPTED
C
  700 CALL LINE(1,NPRE,NPRE,T,Y,ALPHA,BETA)
      DIFS=0
      DO 718 I=1,NPRE
      DIFI=Y(I)-ALPHA-(BETA*T(I))
00718 DIFS=DIFS +(DIFI*DIFI)
      YO=ALPHA +(BETA*(PRE+1.0))
      DIFY=ABSF(Y(NPRE+1)-YO)
      TBAR=SUMT/PRE
      YBAR=SUMY/PRE
      TOB=(PRE+1.0)-TBAR
      TOB2=TOB * TOB
      TIB2=0
      DO 728 I=1,NPRE
      TIBI=T(I)-TBAR
00728 TIB2=TIB2 +(TIBI*TIBI)
      CON1= ABSF(((PRE+1.0)/PRE)+(TOB2/TIB2))
      CON1=SQRTF(CON1)
      CON2= ABSF(DIFS/(PRE-2.0))
      CON2=SQRTF(CON2)
      SE=CON1*CON2
      TMOOD=DIFY/SE
      MDF=NPRE-2
      VARM=DIFS/PRE
      IF(MORE.EQ.1)413,417
00413 WRITE OUTPUT TAPE 6,56,TMOOD,MDF
      WRITE OUTPUT TAPE 6,57,SE,YO,Y(NPRE1)
      WRITE OUTPUT TAPE 6,58,BETA,ALPHA,VARM
      GO TO 418
00417 WRITE OUTPUT TAPE 6,56,TMOOD,MDF
  418 IF(LEV.NE.0)900,429
C
C       WALKER-LEV CALCULATIONS,IF OPTED
C
  900 SUMT1=0
      SUMY1=0
      SUMTY1=0
      SUMTT1=0
      SUMYY1=0
      DO 910 I=1,NPRE
      SUMT1=SUMT1+T(I)
      SUMY1=SUMY1+Y(I)
      SUMTY1=SUMTY1+(T(I)*Y(I))
      SUMTT1=SUMTT1+(T(I)*T(I))
00910 SUMYY1=SUMYY1+(Y(I)*Y(I))
      SUMT2=0
      SUMY2=0
      SUMTT2=0
      SUMYY2=0
      DO 921 I=NPRE1,NTOT
      SUMT2=SUMT2+T(I)
      SUMY2=SUMY2+Y(I)
      SUMTY2=SUMTY2 +(T(I)*Y(I))
      SUMTT2=SUMTT2 +(T(I)*T(I))
00921 SUMYY2=SUMYY2+(Y(I)*Y(I))
      CTY1=SUMTY1-((SUMT1*SUMY1)/PRE)
      CTY2=SUMTY2-((SUMT2*SUMY2)/POST)
      CTT1=SUMTT1-((SUMT1*SUMT1)/PRE)
      CTT2=SUMTT2-((SUMT2*SUMT2)/POST)
      CYY1=SUMYY1-((SUMY1*SUMY1)/PRE)
      CYY2=SUMYY2-((SUMY2*SUMY2)/POST)
      CTYW=CTY1+CTY2
      CTTW=CTT1+CTT2
      CYYW=CYY1+CYY2
      CTYT=(SUMTY1+SUMTY2)-(((SUMT1+SUMT2)*(SUMY1+SUMY2))/TOT)
      CTTT=(SUMTT1+SUMTT2)-(((SUMT1+SUMT2)*(SUMT1+SUMT2))/TOT)
      CYYT=(SUMYY1+SUMYY2)-(((SUMY1+SUMY2)*(SUMY1+SUMY2))/TOT)
      CYYB=CYYT-CYYW
      C2TYW=CTYW*CTYW
      C2TYT=CTYT*CTYT
      C2TY1=CTY1*CTY1
      C2TY2=CTY2*CTY2
      S1=((C2TY1/CTT1)+(C2TY2/CTT2))-(C2TYW/CTTW)
      S2=CYYW-((C2TY1/CTT1)+(C2TY2/CTT2))
      FLEV1=(S1/S2)*(TOT-4.0)
      L1DF=NTOT-4
      SSD=(CTTW*CYYW)-C2TYW
      FLEV2=(C2TYW/SSD)*(TOT-4.0)
      L2DF=L1DF
      SB=CYYB+(C2TYW/CTTW)-(C2TYT/CTTT)
      SW=S1+S2
      FLEV3=(SB/SW)*(TOT-3.0)
      L3DF=NTOT-3
      IF(MORE.NE.0)954,423
00954 BS1=CTY1/CTT1
      Y1BAR=SUMY1/PRE
      T1BAR=SUMT1/PRE
      AS1=Y1BAR-(BS1*T1BAR)
      BS2=CTY2/CTT2
      Y2BAR=SUMY2/POST
      T2BAR=SUMT2/POST
      AS2=Y2BAR-(BS2*T2BAR)
      BW=CTYW/CTTW
      AW1=Y1BAR-(BW*T1BAR)
      AW2=Y2BAR-(BW*T2BAR)
      BT=CTYT/CTTT
      AT=((SUMY1+SUMY2)/TOT)-(BT*((SUMT1+SUMT2)/TOT))
      TX=PRE+0.5
      YTX1=(BS1*TX)+AS1
      YTX2=(BS2*TX)+AS2
      YTXC1=(BW*TX)+AW1
      YTXC2=(BW*TX)+AW2
      YTXT=(BT*TX)+AT
```

```
00422 WRITE OUTPUT TAPE 6,59
      WRITE OUTPUT TAPE 6,60,FLEV1,L1DF
      WRITE OUTPUT TAPE 6,61,S1,S2
      WRITE OUTPUT TAPE 6,62
      WRITE OUTPUT TAPE 6,63,BS1,AS1
      WRITE OUTPUT TAPE 6,78,Y1X1
      WRITE OUTPUT TAPE 6,64,BS2,AS2
      WRITE OUTPUT TAPE 6,78,Y1X2
      WRITE OUTPUT TAPE 6,65
      WRITE OUTPUT TAPE 6,63,BW,AW1
      WRITE OUTPUT TAPE 6,78,Y1XC1
      WRITE OUTPUT TAPE 6,64,BW,AW2
      WRITE OUTPUT TAPE 6,78,Y1XC2
      WRITE OUTPUT TAPE 6,66,FLEV2,L2DF
      WRITE OUTPUT TAPE 6,136,C2TYW,SSD
      WRITE OUTPUT TAPE 6,67,FLEV3,L3DF
      WRITE OUTPUT TAPE 6,61,SB,SW
      WRITE OUTPUT TAPE 6,68
      WRITE OUTPUT TAPE 6,69,BT,AT
      WRITE OUTPUT TAPE 6,78,Y1XT
      GO TO 429
  423 WRITE(6,59)
      WRITE OUTPUT TAPE 6,60,FLEV1,L1DF
      WRITE OUTPUT TAPE 6,66,FLEV2,L2DF
      WRITE OUTPUT TAPE 6,67,FLEV3,L3DF
  429 IF(MOOD2.NE.0)1200,473
C
C       DOUBLE MOOD CALCULATIONS, IF OPTED
C
 1200 CALL LINE(1,NPRE,NPRE,T,Y,ALPHA1,BETA1)
      Y01=ALPHA1 + (BETA1 * (PRE + 0.5))
      CALL LINE(NPRE1,NTOT,NPOST,T,Y,ALPHA2,BETA2)
      Y02 =ALPHA2 +  (BETA2*(PRE+0.5))
      SUMY=0$SUMT=0$SUMTT=0$SUMYY=0$SUMTY=0
      DO 1210 I=1,NPRE
      SUMY =SUMY  + Y(I)
      SUMT =SUMT  + T(I)
      SUMTT =SUMTT  + (T(I)*T(I))
      SUMYY=SUMYY+(Y(I)*Y(I))
 1210 SUMTY=SUMTY+(T(I)*Y(I.)
      CYY1=SUMYY - ((SUMY*SUMY)/PRE)
      CTT1=SUMTT - ((SUMT*SUMT)/PRE)
      CTY1=SUMTY - ((SUMT*SUMY)/PRE)
      TBAR1=SUMT/PRE
      SUMY=0$SUMT=0$SUMTT=0$SUMYY=0$SUMTY=0
      DO 1234 I=NPRE1,NTOT
      SUMY= SUMY + Y(I)
      SUMT= SUMT + T(I)
      SUMTT=SUMTT + (T(I)* T(I))
      SUMYY=SUMYY + (Y(I)* Y(I))
 1234 SUMTY=SUMTY + (T(I)* Y(I))
      CYY2=SUMYY-((SUMY*SUMY)/POST)
      CTT2=SUMTT-((SUMT*SUMT)/POST)
      CTY2=SUMTY-((SUMT*SUMY)/POST)
      TBAR2=SUMT/POST
      DIFY= ABSF(Y01-Y02)
      C2TY1=CTY1*CTY1
      U1=CYY1 -(C2TY1/CTT1)
      V1=(((PRE+ 0.5)-TBAR1)*((PRE+0.5)-TBAR1))/CTT1
      C2TY2=CTY2 *CTY2
      U2= CYY2 -(C2TY2/CTT2)
      V2=(((PRE+0.5) -TBAR2)*((PRE+0.5)-TBAR2))/CTT2
      TOP=(U1 + U2)*((TOT/(PRE*POST))+ V1 +V2)
      SE=  ABSF(TOP/(PRE + POST-4.0))
      SE=SQRTF(SE)
      TMOOD2=DIFY/SE
      M2DF=NTOT-4
      IF(MORE.NE.0)1257,433
 1257 DIFSQ =0
      DO 1260 I=1,NPRE
      DIFI  =Y(I)-ALPHA1 -(BETA1 *T(I))
 1260 DIFSQ=DIFSQ+(DIFI*DIFI)
      VAR1=DIFSQ/PRE
      DIFSQ=0
      DO 1265 I=NPRE1 ,NTOT
      DIFI  = Y(I)-ALPHA2-(BETA2 * T(I))
 1265 DIFSQ=DIFSQ+(DIFI*DIFI)
      VAR2=DIFSQ/POST
      WRITE OUTPUT TAPE 6,73,TMOOD2,M2DF
      WRITE OUTPUT TAPE 6,74,SE
      WRITE OUTPUT TAPE 6,75,BETA1,ALPHA1
      WRITE OUTPUT TAPE 6,76,VAR1,Y01
      WRITE OUTPUT TAPE 6,77,BETA2,ALPHA2
      WRITE OUTPUT TAPE 6,76,VAR2,Y02
      GO TO 473
  433 WRITE OUTPUT TAPE 6,73,TMOOD2,M2DF
  473 IF(NPLOT.NE.0)476,100
  476 WRITE(6,29)
      WRITE(6,30) ITITLE
C
C        DATA PLOT, IF OPTED
C
 2500 BCD1=1HO
      BCD2=1HX
      IF(NPLOT-1)2532,2531
 2531 IF(NTOT-50)2501,2501,2502
 2501 CALL STPLT2(1,T(1),Y(1),NTOT)
      CALL PLOT3(BCD1,T(1),Y(1),NPRE)
      CALL PLOT3(BCD2,T(NPRE1),Y(NPRE1),NPOST)
      GO TO 2525
 2502 IF(NPRE-25) 2511,2511,2512
 2511 CALL STPLT2(1,T(1),Y(1),50)
      CALL PLOT3(BCD1,T(1),Y(1),NPRE)
      NUM2= 50-NPRE
      CALL PLOT3(BCD2,T(NPRE1),Y(NPRE1),NUM2)
      GO TO 2525
 2512 IF(NPOST-25) 2513,2513,2514
 2513 NUM1=50-NPOST
      I1=1+(NTOT-50)
      CALL STPLT2(1,T(I1),Y(I1),50)
      CALL PLOT3(BCD1,T(I1),Y(I1),NUM1)
      CALL PLOT3(BCD2,T(NPRE1),Y(NPRE1),NPOST)
      GO TO 2525
 2514 I1=1+(NPRE-25)
```

```
      CALL STPLT2(1,T(I1),Y(I1),50)
      CALL PLOT3(BCD1,T(I1),Y(I1),25)
      CALL PLOT3(BCD2,T(NPRE1),Y(NPRE1),25)
      GO TO 2525
 2525 WRITE OUTPUT TAPE 6,147
      CALL FPLOT4(30,30H                           OCCASION VALUES)
      WRITE OUTPUT TAPE 6,149
      GO TO 100
 2532 CALL LINE(1,NTOT,NTOT,T,Y,BTT,SLOPET)
      DO 2551 I=1,NTOT
 2551 ZTOT(I)=(SLOPET * T(I))+ BTT
      CALL LINE(1,NPRE,NPRE,T,Y,B1,SLOPE1)
      DO 2553 I=1,NPRE
 2553 ZPART(I)=(SLOPE1 * T(I))+ B1
      CALL LINE(NPRE1,NTOT,NPOST,T,Y,B2,SLOPE2)
      DO 2555 I=NPRE1,NTOT
 2555 ZPART(I)=(SLOPE2 * T(I))+B2
      BCDP1=1H,
      BCDP2=1H.
      BCDPT=1H*
      IF(NTOT-50)2561,2561,2562
 2561 NUM1=NPRE
      NUM2=NPOST
      NUMT=NTOT
      IB=1
      IN=NTOT
      GO TO 2585
 2562 IF(NPRE-25)2571,2571,2572
 2571 NUM1=NPRE
      NUM2=50-NPRE
      NUMT=50
      IB=1
      IN=50
      GO TO 2585
 2572 IF(NPOST-25)2573,2573,2574
 2573 NUM1=50-NPOST
      NUM2=NPOST
      NUMT=50
      IB=1+(NTOT-50)
      IN=IB+49
      GO TO 2585
 2574 NUM1=25
      NUM2=25
      NUMT=50
      IB=1+(NPRE-25)
      IN=IB+49
      GO TO 2585
 2585 CALL STPLT2(I,T(IB),Y(IB),NUMT)
      CALL PLOT3(BCDP1,T(IB),ZPART(IB),1)
      CALL PLOT3(BCDP1,T(NPRE),ZPART(NPRE),1)
      CALL PLOT3(BCDP2,T(NPRE1),ZPART(NPRE1),1)
      CALL PLOT3(BCDP2,T(IN),ZPART(IN),1)
      CALL PLOT3(BCDPT,T(IB),ZTOT(IB),1)
      CALL PLOT3(BCDPT,T(IN),ZTOT(IN),1)
      DO 2586 I=IB,NPRE
 2586 CALL PLOT3(BCD1,T(I),Y(I),NUM1)
      DO 2587 I=NPRE1,IN
 2587 CALL PLOT3(HCD2,T(I),Y(I),NUM2)
      WRITE OUTPUT TAPE6,147
      WRITE OUTPUT TAPE 6,148
      CALL FPLOT4(30,30H                   OCCASION VALUES)
      WRITE OUTPUT TAPE 6,149
      GO TO 100
   91 WRITE(6,991)
  999 STOP
C
C       FORMAT STATEMENTS
C
   10 FORMAT(10A8)
   11 FORMAT(3I3)
   12 FORMAT(6I1)
   29 FORMAT(1H1)
   30 FORMAT(1H ,10A8)
  120 FORMAT(10F10.2)
  130 FORMAT(33H0PRE-EXPERIMENTAL OCCASION VALUES)
  131 FORMAT(34H POST-EXPERIMENTAL OCCASION VALUES)
   54 FORMAT(32H0SERIAL CORRELATION COEFFICIENTS)
   55 FORMAT( 6H R(1)=F6.2,8H   R(2)=F6.2,8H   R(3)=F6.2,8H   R(4)=F6.2)
   45 FORMAT(1H ,5X,31HSLOPE AND INTERCEPT NOT REMOVED)
   46 FORMAT(1H ,5X,55HSLOPE AND INTERCEPT REMOVED (TOTAL REGRESSION EST
     1IMATE))
   47 FORMAT(1H ,5X,72HSLOPE AND INTERCEPT REMOVED (SEPARATE PRE AND POS
     1T REGRESSION ESTIMATES))
   56 FORMAT(23H0MOOD TEST             T=F7.2,7H    DF=I3)
   57 FORMAT(19H    STANDARD ERROR=F12.2,13H  Y ESTIMATE=F12.2,4H  Y=F12
     1.2)
   58 FORMAT(10H    SLOPE=F12.3,13H   INTERCEPT=F12.3,12H   VARIANCE=F15
     1.2)
   59 FORMAT(17H0WALKER-LEV TESTS)
   60 FORMAT(23H    TEST 1              F=F7.2,9H    DF=1,I3)
   61 FORMAT(19H       NUMERATOR SS=F15.2,19H    DENOMINATOR SS=F15.2)
   62 FORMAT(41H       SEPARATE GROUP REGRESSION ESTIMATES)
   63 FORMAT(30H          PRE-X GROUP    SLOPE=F12.3,13H   INTERCEPT=F12.
     13)
   64 FORMAT(30H          POST-X GROUP   SLOPE=F12.3,13H   INTERCEPT=F12.
     13)
   65 FORMAT(47H       COMMON WITHIN GROUPS REGRESSION ESTIMATES)
   66 FORMAT(23H    TEST 2              F=F7.2,9H    DF=1,I3)
   67 FORMAT(23H    TEST 3              F=F7.2,9H    DF=1,I3)
   68 FORMAT(41H       COMBINED GROUPS REGRESSION ESTIMATE)
   69 FORMAT(30H          TOTAL GROUP    SLOPE=F12.3,13H   INTERCEPT=F12.
     13)
   73 FORMAT(23H0DOUBLE MOOD TEST    T=F7.2,7H    DF=I3)
   74 FORMAT(19H    STANDARD ERROR=F12.2)
   75 FORMAT(25H    PRE-X GROUP    SLOPE=F12.3, 13H   INTERCEPT=F12.3)
   76 FORMAT(22H              VARIANCE=F15.2,21H   ESTIMATED VALUE Y=F12.
     12)
   77 FORMAT(25H    POST-X GROUP   SLOPE=F12.3,13H   INTERCEPT=F12.3)
   78 FORMAT(48H                           ESTIMATED Y VALUE AT X=F12.2)
  136 FORMAT(19H       NUMERATOR SS=F15.2,19H    DENOMINATOR SS=F15.2)
  147 FORMAT(52H PLOTTING CHARACTERS. PRE-CHANGE(O), POST-CHANGE(X).)
  148 FORMAT(98H PLOTTING CHARACTERS-POINTS BASED ON REGRESSION ESTIMATE
     1S. PRE-CHANGE(,), POST-CHANGE(.), TOTAL(*))
  149 FORMAT(72H
     1   I14* INTERVALS)
 991 FORMAT(1H ,*ERROR IN SPECIFICATION CARD. PROGRAM TERMINATED*)
      END
      SUBROUTINE SERIAL(NT,N,V,RU)
      DIMENSION V(400)
      SUMXY=0
      SUMX=0
      SUMY=0
      SUMXX=0
      SUMYY=0
      NN=NT-N
      DO 6002 I=1,NN
      IN=I+N
      SUMXY=SUMXY+(V(I)*V(IN))
      SUMX=SUMX+V(I)
      SUMY=SUMY+V(IN)
      SUMXX=SUMXX+(V(I)*V(I))
 6002 SUMYY=SUMYY+(V(IN)*V(IN))
      TOT=NN
      TOP=(TOT * SUMXY)-(SUMX * SUMY)
      DENOM1= ABSF((TOT*SUMXX)-(SUMX * SUMX))
      DENOM1=SQRTF(DENOM1)
      DENOM2= ABSF((TOT *SUMYY)-(SUMY * SUMY))
      DENOM2=SQRTF(DENOM2)
      RU=TOP/(DENOM1*DENOM2)
      RETURN
      END
      SUBROUTINE LINE(N1,N2,NT,T,Y,ALPHA,BETA)
      DIMENSION T(400),Y(400)
      SUMY=0
      SUMT=0
      SUMTT=0
      SUMTY=0
      DO 6001 I=N1,N2
      SUMTY=SUMTY+(T(I)*Y(I))
      SUMT=SUMT + T(I)
      SUMY=SUMY + Y(I)
 6001 SUMTT= SUMTT + (T(I)*T(I))
      TOT=NT
      TOP=(TOT*SUMTY)-(SUMT*SUMY)
      DENOM=(TOT*SUMTT)-(SUMT * SUMT)
      BETA=TOP/DENOM
      TOP=(SUMTT * SUMY)-(SUMT * SUMTY)
      ALPHA=TOP/DENOM
      RETURN
      END
```

```
TEST DATA FOR TIME SERIES PROGRAM

PRE-EXPERIMENTAL OCCASION VALUES
     9.10      9.58      8.73      9.62      9.20      9.10      9.10      9.53     10.00      9.13
     9.62     10.30     10.30     10.38      9.61      9.69      9.52
POST-EXPERIMENTAL OCCASION VALUES
     9.72     10.40     10.98     10.80     11.23     10.90     10.70     10.80     10.13     11.60
    10.43     11.42

SERIAL CORRELATION COEFFICIENTS
     SLOPE AND INTERCEPT REMOVED (TOTAL REGRESSION ESTIMATE)
R(1)=  0.02   R(2)=  0.14   R(3)= -0.26   R(4)= -0.40
     SLOPE AND INTERCEPT REMOVED (SEPARATE PRE AND POST REGRESSION ESTIMATES)
R(1)= -0.04   R(2)=  0.13   R(3)= -0.24   R(4)= -0.39
     SLOPE AND INTERCEPT NOT REMOVED
R(1)=  0.67   R(2)=  0.72   R(3)=  0.54   R(4)=  0.53

MOOD TEST            T=   0.71    DF= 15
   STANDARD ERROR=         0.45  Y ESTIMATE=       10.04  Y=         9.72
   SLOPE=        0.054   INTERCEPT=        9.077   VARIANCE=           0.14

WALKER-LEV TESTS
   TEST 1            F=   0.03    DF=1, 25
     NUMERATOR SS=          0.01    DENOMINATOR SS=           4.96
     SEPARATE GROUP REGRESSION ESTIMATES
        PRE-X GROUP    SLOPE=        0.054   INTERCEPT=        9.077
                        ESTIMATED Y VALUE AT X=       10.01
        POST-X GROUP   SLOPE=        0.061   INTERCEPT=        9.324
                        ESTIMATED Y VALUE AT X=       10.39
     COMMON WITHIN GROUPS REGRESSION ESTIMATES
        PRE-X GROUP    SLOPE=        0.055   INTERCEPT=        9.061
                        ESTIMATED Y VALUE AT X=       10.03
        POST-X GROUP   SLOPE=        0.055   INTERCEPT=        9.449
                        ESTIMATED Y VALUE AT X=       10.42
   TEST 2            F=   8.52    DF=1, 25
     NUMERATOR SS=        932.39    DENOMINATOR SS=        2737.16
   TEST 3            F=   1.50    DF=1, 26
     NUMERATOR SS=          0.29    DENOMINATOR SS=           4.97
     COMBINED GROUPS REGRESSION ESTIMATE
        TOTAL GROUP    SLOPE=        0.075   INTERCEPT=        8.929
                        ESTIMATED Y VALUE AT X=       10.24

DOUBLE MOOD TEST     T=   1.10    DF= 25
   STANDARD ERROR=         0.34
   PRE-X GROUP     SLOPE=          0.054   INTERCEPT=        9.077
             VARIANCE=             0.14   ESTIMATED VALUE Y=       10.01
   POST-X GROUP    SLOPE=          0.061   INTERCEPT=        9.324
             VARIANCE=             0.21   ESTIMATED VALUE Y=       10.39
```

TEST DATA FOR TIME SERIES PROGRAM
PLOTTING CHARACTERS. PRE-CHANGE(O), POST-CHANGE(X).
PLOTTING CHARACTERS-POINTS BASED ON REGRESSION ESTIMATES. PRE-CHANGE(•), POST-CHANGE(.), TOTAL(*)



References

Campbell, D. T. From description to experimentation: interpreting trends as quasi-experiments. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wis.: Univer. Wisconsin Press, 1963, pp. 212-242

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of Research on teaching. Chicago: Rand McNally, 1963, pp. 171-246.

Grenander, U., & Rosenblat, M. Statistical analysis of stationary time series. New York: Wiley, 1957.

Holtzman, Wayne, H. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wis.: Univer. Wisconsin Press, 1963, pp. 199-211.

Mood, S. M. Introduction to the theory of statistics. New York: McGraw-Hill, 1950

Sween, J., and Campbell, D. T. A study of the effect of proximally autocorrelated error on tests of significance for the interrupted time series quasi-experimental design. Mimeographed research report, Department of Psychology, Northwestern University, 1965.

Walker, H. M., & Lev, J. Statistical Inference. New York: Holt, 1953.

Appendix I

Interpretation of Data Plot with Missing Regression Points

If a 2 is placed in Col. 6 of the Test Specification Card, initial and terminal points for the regression lines based on pre-change and post-change values and on the total series are indicated on the data plot. The printing characters for the regression estimates are as follows:

, pre-change regression estimates
. post-change regression estimates
* total series regression estimates

It may sometimes occur that some of these points are missing from the plot. This is because the RUMPLOT subroutine prints only one point at a particular co-ordinate location. If more than one point falls on the same location, one point is printed and the others are surpressed. It is usually possible to determine the location of the missing point directly, although in some instances additional information regarding slopes and intercepts is required. The outline below will facilitate the location of missing points.

(1) Missing total regression estimates (*)

The initial point for the total series regression line is located at time t_1. If there are more than 25 pre-change points it is located at time t_{N_1-24}. If the initial point for the total regression line is not plotted it is located at the first plotted pre-change occasion value (0).

The terminal point for the total series regression line is located at time t_{N_2}. If there are more than 25 post-change points it is located at time t_{N_1+25}. If the terminal point for the total regression line is not plotted it is located at the last plotted post-change occasion value (X).

(2) Missing pre-change regression estimates (,)

The initial point for the pre-change regression line is located at time t_1.

If there are more than 25 pre-change points, it is located at time t_{N_1-24}. If the initial pre-change regression estimate is not plotted _and_ if the initial total regression estimate is not plotted, the initial pre-change regression estimate is located at the first plotted occasion value (0). However, if the initial total regression estimate _is_ plotted, additional information is needed to locate the missing point (it may be at * or 0). Namely, the location of the missing point must be determined from the slope and intercept of the regression line.

The _terminal_ _point_ for the pre-change regression line is located at time t_{N_1}. If this point is missing it is located at the last pre-change occasion value (0).

(3) _Missing post-change regression estimates (.)_

The _initial_ _point_ for the post-change regression line is located at time t_{N_1+1}. If this point is missing, it is located at the first post-change occasion value (X).

The _terminal_ _point_ for the post-change regression line is located at time t_{N_2}. If there are more than 25 post-change points, the terminal point is located at time t_{N_1+25}. If the terminal point for the post-change regression line is not plotted _and_ if the terminal point for the total regression line is not plotted, the missing post-change regression estimate is located at the last plotted occasion value (X). However, if the total terminal regression estimate _is_ plotted, additional information is needed to locate the missing point (it may be at X or *). Namely, the correct location for the missing point must be determined from the slope and intercept of the regression line.

Appendix II

Program Modifications for Fortran II and IV

Program TIMEX is written in CDC 3400 FORTRAN for use with card input and printed output. Modifications for FORTRAN II and IV for the IBM 709/7090 are indicated by reference to statement number. When a statement has no number, modifications are indicated by the number of statements "plus" or "minus" the nearest statement number which are needed to locate the modification.

Statement Number	Modification
100 minus 2	DIMENSION statement should read (ITITLE(12), IFMDAT(12) Also change Card Preparation instructions to indicate that Title Card and Variable Format Card should be punched only in Col. 1-72.
100 minus 1	TYPE statement not required.
1100 minus 1	Change to Arithmetic IF statement.
700 minus 1	Change to Arithmetic IF statement.
418	Change to Arithmetic IF statement.
429	Change to Arithmetic IF statement.
1200 plus 4	$ notation not allowed. Change to separate statements.
1210 plus 5	$ notation not allowed. Change to separate statements.
1257 minus 1	Change to Arithmetic IF statement.
473	Change to Arithmetic IF statement.
991	* notation not allowed. Change to H-specification.

In addition, check all READ and WRITE statements for compatibility with FORTRAN II and IV.

I

Supplementary Note:

Since the research reported herein was done, a highly relevant test of significance has been reported:

> G. E. P. Box and George C. Tiao, "A change in level of a non-stationary time series," Biometrika, 1965, 52, 181-192.

In particular, the Box and Tiao approach avoids the assumption of linearity, in exchange for other, probably more reasonable, assumptions. What is needed is a computer program for the exact distribution computation by numerical methods for their test when δ and γ_o are unknown (pp. 189-191), plus a testing of the formula against null models such as those used here, plus a testing of our linear formulas against the null data generated according to the Box and Tiao assumptions.

1

Time Series Data in the Quasi-Experimental Analysis of the Connecticut Speeding Crackdown*

H. Lawrence Ross, New York University
and
Donald T. Campbell, Northwestern University

On December 23, 1955 Governor Ribicoff initiated an intensive crackdown upon speeding in the State of Connecticut. The change in enforcement practices was so abrupt and strong as to make plausible an analysis of time series data of highway deaths and other measures for evidence of effects. The purposes of this report are two-fold: on the one hand to report the results of the Connecticut experiment, on the other to use this setting for a methodological exposition of the possibilities and problems in using time series data extracted from public records as quasi-experiments. The report will start with a description of the speeding crackdown, and then move into the quasi-experimental analysis.

Resume of the Connecticut Crackdown on Speeding

In 1955, in the State of Connecticut, 324 people were killed in automobile accidents. The death rate per 100,000 population in motor vehicle accidents had reached a record high for the decade of the fifties as the usually hazardous Christmas holidays approached. Two days before Christmas, Governor Ribicoff of Connecticut initiated an unparalleled attempt to control traffic deaths by political means, and announced his crack-down on speeders on the highways of that state.

Ribicoff reasoned that excess speed was the most common contributing

---

*The preparation of this paper was supported in part by U.S. Office of Education, Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

factor to traffic fatalities, and that control of speed would result in diminished fatalities. He believed that previous efforts to control speeding under the usual court procedure and under the existing point system were inadequate. In a study of three months' records of the police court in Hartford, it was noted that no more than half the persons originally charged with speeding were so prosecuted, with the charge often diminished to less serious violations. Ribicoff wanted to initiate a program with reliable procedures and with strong sanctions as a means to control speeding.

On December 23, 1955, Governor Ribicoff announced that in the future all persons convicted of speeding would have their licenses suspended for thirty days at the first offense. A second violation was to mean a sixty-day suspension, and a third violation for speeding would result in indefinite suspension of the driver's license, subject to a hearing after ninety days.

The decree was put into force through the Governor's appointive power over local judges. Under Connecticut practice, the Motor Vehicle Department was suspending licenses on the recommendation of police court judges. The judges were appointed by the Governor, who threatened loss of reappointment in 1957 to judges who appeared lax in the conviction of speeders, or who did not recommend suspension of licenses to the Motor Vehicle Department.

In the first three months of 1956, suspensions for speeding numbered 2855, an increase of almost 2700 over the corresponding period in 1955. There were ten fewer fatalities, and 765 fewer arrests for speeding. The Governor was reported "encouraged" by the drop in violations and in fatalities. The press quoted him as saying, "This is positive proof that operators are not only driving slower, but are driving better."

By the end of May, deaths had declined from 122 in 1955 to 107 in 1956.

Suspensions for speed numbered 4559, as against 209 in 1955. Arrests for speeding had dropped 53%. The Governor received a telegram of commendation for the program from the National Safety Council.

At the end of June there were 22 fewer fatalities than in the first six months of 1955, representing a 15% reduction. Suspensions for speed in the first six months of the year had risen from 231 to 5398, and arrests had declined from 4377 to 2735. Ribicoff announced: "Connecticut has succeeded in stopping the upward surge in highway deaths, and in the first six months of this year, contrary to the national trend, we have saved lives. Fewer people died on the highways this year than in the same period last year, in Connecticut. We did it by enforcing the law, something the safety experts said couldn't be done because the people wouldn't be behind it."

In July a new state police program, using unmarked police cars and making extensive use of radar, was inaugurated. The police issued a report stating that 2% of the cars observed by radar on July 4 were found to be speeding; at a later date, no speeders were found among 53,000 cars similarly observed.

Late summer produced a very high number of traffic fatalities in Connecticut. By the beginning of September, 194 people had been killed in the year, almost equal to the 195 in the comparable period of the previous year. The accident "epidemic" was embarrassing to the authorities, who defended the speed crackdown on the grounds that (a) the fatality rate remained low in comparison to the national trend, which showed a seven per cent increase; (b) the exposure to accidents in the state had increased by 100 million vehicle miles; and (c) that the total accident rate had risen, so the proportion of fatal accidents to total accidents had declined.

Fatalities were fewer in the Fall of 1956, and by the end of the year Connecticut could count 284 deaths in traffic as against 324 in 1955. The Governor stated, "With the saving of forty lives in 1955, a reduction of 12.3 per cent from the 1955 motor vehicle death toll--we can say the program is definitely worthwhile."

The crackdown on speeders is still in effect in Connecticut, although it is no longer the subject of newsworthy comment. It was not completely beneficial politically for the Democratic Governor. From the start, there were problems with neighboring states, originating a relatively large share of Connecticut traffic, which at first refused to suspend licenses of drivers convicted of speeding in Connecticut, although today reciprocity has been established. More important, many powerful groups, as well as particular individuals, within the State resented the direct effects of the crackdown. Members of the Republican Party wanted the program "tempered with justice." The Teamsters sponsored a bill to eliminate compulsory suspension of the license on a first offense, and other legislation, granting restricted driving permits for "hardship" cases was introduced. These efforts were not successful in officially moderating the crackdown policy. The people of Connecticut and their officials are paying what in many instances appears a stiff price for this program. Few will feel the price is too high if it can be shown that as many as forty lives per year are being sveed in Connecticut. However, the question may be raised as to what extent the results claimed for the program in 1956 are valid, and what conclusions may be drawn from a longer period of observation.

Quasi-experimental Analysis: Before and After Measures

Figure 1 portrays graphically the final results for 1956, compared with 1955, and expressed as the rate per 100,000,000 miles of driving. Properly prepared, such results can look impressive. The decline might well be statistically significant, expressed as a proportion of the estimated number of drivers on the highway in each year.

The evidence presented can be considered as a kind of experiment: there is a pretest (the 1955 figures) an experimental treatment, and a posttest. A substantial change is noted, which one would like to ascribe to the experimental treatment. It is in the spirit of the program of quasi-experimental analysis[1] of field data to encourage such interpretation, provided consideration is given to the plausible rival hypotheses left uncontrolled with supplementary analyses being added to eliminate these where possible. In this tradition of analysis, the data of Figure 1 constitute a One-Group Pretest-Posttest Design. As such it fails to control for five common threats to the validity of experiments.

(1) History, i.e., other events occurring between pretest and posttest which might account for the change. Thus 1956 might have been a particularly dry year, with fewer accidents due to rain, snow, and ice. Thus 1956 might have been--though we know it is not so in this case--a wartime year with gas rationing and fewer, slower, drivers.

(2) Maturation, which applied to the present situation can be used to cover autonomous long term secular changes differing from history in not being related to specific events. For example, one can classify here the general long term trend for a reduction of mileage-based traffic deaths, presumably due to an increasing availability of through-ways.

(3) Testing--the effect of a pretest in introducing a change which would have occurred anyway even without the experimental treatment. In this case, testing is not a very plausible rival hypothesis since the pretest had been going on for years. Conceivably, however, the measurement of the traffic death rate and the publicizing of it could change driver caution in a subsequent year.

(4) Instrumentation, a shifting the measuring instrument independent of any change in the phenomenon measured. This too is inplausible in the present instance, although a shift in the government agency recording the fatality statistics could account for a shift, for note the dramatic 20% increase in suicides in Prussia between 1882 and 1883, when record keeping was transferred from local police to national civil service.[2] Similarly, Orlando Wilson's reforms of the police system in Chicago led to dramatic increases in rates for most crimes, due entirely, he feels, to more complete recording.[3] In the present setting, note that the mileage based mortality rate is computed by using the gallons of gasoline sold in the state to estimate miles driven. A revision in the miles-per-gallon estimate figure could cause such a shift. A decrease in the actual miles obtained per gallon, as through engines of larger horsepower or driving at higher speeds, could masquerade as a lowered mileage-based fatality rate, through inflating the estimate of miles driven.

(5) Regression. Where the group has been selected for treatment just because of its extreme performance on the pretest, it follows that if pretest and posttest are imperfectly correlated, on the average the posttest will be less extreme than the pretest. This regression is a tautological restatement of the imperfect correlation and the pretest extremity, the *r* in the correlation

coefficient standing for the percentage of regression toward the mean. Such regression can be seen as plausibly operating here in two dimensions: of all states in 1955, this treatment was most likely to be applied to one with an exceptionally high traffic casualty rate; or of all years subsequent to which such a treatment was apt to be applied, the most likely is a year in which traffic fatalities are exceptionally high.

In the true experiment, the treatment is assigned randomly, without relation to the prior state of the dependent variable. For these, the correlation between pretest scores and treatment-not-treatment is zero. The same is true in the most interpretable of quasi-experiments: the treatment is applied with no plausible relationship to the prior status of the group. Thus an analysis of the effects of a tornado or an earthquake can be made in confidence that the pretreatment measures did not cause the tornado or earthquake. Not so here: the high 1955 scores plausibly caused the treatment. That 1965 was less extreme is expected because of regression.[4]

Interrupted Time-Series Analysis

Figure 2 plots fatality rates for five years before and four years after the crackdown. This mode of quasi-experimental analysis has been labeled "Interrupted Time-Series"[5] to point to a distinction between it and the time-series analyses of economics. In the economic analysis, the exogenous variable for which causal imputation is examined, is a continuously present variable, occurring in differing degrees. In the interrupted time-series, the "causal" variable is examined as an event or change occurring at a specified time, and typically only once.

The Interrupted Time-Series design represents a more extended use of data which are often available when only before and after measures are reported. It has the advantage of making possible the examination of the hypotheses of maturation and regression, plus an examination of the "normal" level of temporal instability. (For note that even though the 1955-56 shift were highly statistically significant, it would not be interpretable if there were frequent other "inexplicable" but equally significant changes.)

Looking at the time series data in Figure 2 destroys most of the impression of significant effect. The most exceptional point is 1955, not 1956. The largest shift in the curve is 1954-55, not 1955-56. Causality from pretest to treatment looks much more plausible than causality from treatment to posttest. There is also in the pretest data some slight indication of a downward slope,--broken up for 1955. Thus with some plausibility, the "effect" might be attributed to an irregular downward trend.

The present writers would agree with those sociological analysts who place tests of significance in the role of auxiliary aids rather than the final arbiters of truth. Nonetheless, the opportunities for wishful perception in the interpretation of trends[6] have argued for an exploration of tests of significance for time-series data. For the moment, only linear hypotheses seem feasibly tested. These focus on comparing the least-squares line of best fit fitted for the pretest points with that for the posttest points. For this, the t for the points at treatment extrapolated from the pretest and posttest is but .17 with 5 degrees of freedom. The F ratio pooling differences in slope and intercept is but .21 with d.f. of 1 and 6.[7] These tests have been found in computer-simulation monte-carlo runs to provide too large F's and t's when the data show high autocorrelation for the residuals when the linear slope has been extracted for adjacent points in time, as compared with more remote points. This is a very common condition in actual social data.[8] This bias is not of importance here. The analyses have been recomputed using crude annual fatalities, with essentially the same outcomes.

One real problem is the very few degrees of freedom. To attempt to overcome this, the analyses have been redone on monthly data. For these, the intercept _t_ becomes but .98 with 104 degrees of freedom, the slope-intercept F is 1.84 with 1 and 105 d.f., where 3.93 would be required for significance at the $p < .05$ level.

Multiple Times Series

In many situations, time-series involving but a single experimental unit will be all that are available, and will be profitably analyzed as a great improvement over the one-measure before and one after study. Some such analyses will be presented below. It is in the spirit of quasi-experimental analyses, however, to make use of all available data that could help to rule out or confirm any plausible rival hypotheses. In a setting such as this, no randomly assigned control group is available. But in quasi-experimentation, even a non-equivalent control group is a help with such plausible rival hypotheses as history, maturation, testing effects, and the like. This is so providing one _does not_ attempt to overcome initial pretest discrepancies by matching--one should instead live with them. For Connecticut, it was judged that a pool of adjacent and similar states (New York, New Jersey, Rhode Island, and Massachusettts) provided a meaningful comparison. Figure 3 plots the annual mileage-base data for the control states along side the Connecticut data. The control data are much smoother, due to the much larger base, i.e., the cancelling out of chance deviations in the annual figures for particular states. The data confirm the secular trend for reduced mileage-based highway fatality rates.

On the other hand, the differences between Connecticut and the control states show a pattern somewhat conforting to the hypothesis that the crackdown made a difference. In the pretest years, Connecticut's rate is rising relative to the others, coming closest in 1955. In the posttest years, Connecticut's rate is steadily dropping faster than are the control states. As the optimal test of these apparent trends, the monthly difference between Connecticut and the control series has been employed in a time-series analysis. For this, the intercepts at treatment show a different in the wrong direction and that $\underline{t}$ is but .63. The F for slope-intercept is but .06, even though visually-graphically it looks more promising than this. We must conclude that there is no evidence whatsoever that the crackdown caused a decrease in fatalities.

Auxiliary Analyses

In this section, we will present data which will both further illustrate time-series analysis and answer a possible question as to whether or not there was a real crackdown.

Figure 4 shows evidence of the crackdown, stated as the percentage of all suspensions of licenses which are for speeding. Unfortunately, for this and the following variables we have been unable to get control-state data, but the one-state time-series is quite convincing in itself. We regard it confirming of the appropriateness of the statistical tests that they too show it significant. The double-extrapolation intercept test provides a $\underline{t}$

of 4.30 which with 6 degrees of freedom (in the computations, but not the graph, 1950 data were used), where 3.71 would be significant at $p < .01$. The slope-intercept F is 12.05, which with 1 and 7 degrees of freedom clearly exceeds the $p < .05$ level of 5.59 and almost reaches the $p < .01$ level of 12.25. The pooled pretest and posttest autocorrelation for lag 1 of departures from the lines of best fit is -.53, showing adjacent points to tend to lie on opposite sides of the linear fit. Such negative values tend to occur in monte-carlo runs with zero true correlation, due to overfitting of obtained data, particularly when few points are involved. It is high positive lag 1 autocorrelations that betoken the overestimates of F and t, so no cautions of interpretation are indicated here, although to be sure, there are too few points to provide a good estimate of correlation. One other test deserves note. Mood[9] provides a t test for the difference of a single observation (e.g., the first posttest measure) from a corresponding value extrapolated from a series (e.g., the pretest series). This t is 4.33 with 4 degrees of freedom, where 3.745 is significant at the $p < .02$ level. This test is particularly resistent in the monte-carlo studies to bias due to autocorrelation.

Figure 5 plots the percentage which speeding violations constitute of all traffic violations. This shows an apparent decline, due presumably to greater conformity to the speed limits rather than to greater enforcement leniency, for note the evidence of severity in Figure 4. While the graphic portrayal is convincing of genuine affect, the statistical tests are not so emphatic. The intercept t is 2.29 with 6 d.f., not significant at the $p < .05$ level. The slope-intercept F is 6.52 with 1 and 7 d.f., exceeding the p 05 level but falling far short of $p < .01$. The Mood test t is 2.66

with 4 d.f., not reaching the $p < .05$ level of 2.78.

From Figure 4, and from the reports cited in the first section of this paper, it is clear that a real change in enforcement behavior resulted. We are convinced that in practice, the proportion of drivers exceeding the speed limit actually decreased. However, over and above these desired effects there were signs of complex unwanted reactions.

Figure 6 shows the percentage of persons arrested while driving with a suspended license, expressed as a percentage of all suspensions. This jumps from an almost consistent zero to some 4 to 6 per cent. Tests of significance confirm the effect. The intercept *t* is 4.30, when with 6 d.f. 3.71 would be significant at $p < .01$. The slope-intercept *F* is 12.05, exactly hitting the $p < .01$ level for 1 and 7 d.f. The Mood *t* reaches an incredible 130.75, due to the very small error term which the very small variance of the pretest scores produces. (While one feels uneasy with a practically zero variance, the consistent pretest zero does genuinely make the later values unlikely.)

Figure 7 shows a reaction on the part of the judicial system. Even with fewer speeding violations reaching the courts (Figure 5), the judges were more lenient in their handling of these cases, the percentage of not guilty decisions going up, presumably due to increased reluctance to impose the drastic penalties involved. Tests of significance are borderline. The intercept *t* is 2.65, which with 6 d.f., is significant at the $p < .05$ level. The slope-intercept F is 8.03, where 5.59 is required for significance at the $p < .05$ level with 1 and 7 d.f. The Mood *t* is 2.42, which with but 4 d.f. fails to reach significance at the $p < .05$ level, for which 2.78 would be required.

Discussion and Conclusions

On the substantive side, the analysis has demonstrated that the Connecticut Speeding Crackdown was a substantial enforcement effort. As to its effects, no appreciable reduction in traffic fatalities has been found, and unwanted side effects have increased as in the illegal behavior of driving without a license and in the greater leniency of the courts. These findings are relevant to applied social science, the sociology of law, and administrative theory, for example.

More important, we believe, is the methodology explored. While the social scientist cannot as a rule experiment on a societal scale, experimentation or abrupt focussed environmental change is continually going on, initiated by government and business. The social scientist adds to his tools for understanding the social system when he attends to such and documents their effects in as thorough a fashion as possible. Insofar as correlational survey approaches differ from experimental analysis, it adds depth to the social scientist's work when he examines the fit of an experimental interpretation with full attention to the uncontrolled competing hypotheses.

The methodology for such quasi-experimental analysis has a long but casual history, and offers much room for development. In such development, it should be remembered that not only are the raw materials shaped by the tools, but in the long run, the tools are shaped by the materials upon which they work. Thus we should not passively accept a methodology as a revealed truth, but rather should test it in use with our materials. Methodology has in fact an empirical history and its constituents have the status of empirical discoveries. Thus the control group design emerged from laboratory experience, and is peculiar to the social sciences and their problems: thus medical research has the placebo control group, and neurophysiology the sham operation control, as achievements of specific research traditions, not as logical dispensations from the philosophy of science or mathematical statistics. So too the methods for quasi-experimentation in settings like the present will emerge from an iteration of effort and criticism, in which many approaches will be rejected.

For example, the tests of significance based upon linear hypotheses represent an obvious oversimplification, a concession to expediency. Such tests can be regarded as but partial checks, although even here a curvilinear trend can masquerade as a significant shift in linear trends. Certainly these tests of significance should not be employed without a graphic presentation that confirms the effect.

Footnotes

*The preparation of this paper was supported in part by U.S. Office of Education, Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

[1] E.g., Donald T. Campbell and Julian S. Stanley, "Experimental and Quasi-experimental Designs for Research on Teaching," in Nathan L. Gage, Ed., *Handbook of Research on Teaching*, Chicago: Rand McNally, 1963, pp. 171-246; Donald T. Campbell, "From Description of Experimentation: Interpreting Trends as Quasi-experiments," in Chester W. Harris, Ed., *Problems in Measuring Change*. Madison: University of Wisconsin Press, 1963; Donald T. Campbell and Keith N. Clayton, "Avoiding Regression Effects in Panel Studies of Communication Impact," *Studies in Public Communication*, No. 3, Chicago: Department of Sociology, University of Chicago, 1961, pp. 99-118, reprinted in Bobbs-Merrill Reprints in Sociology as S-353.

[2] Cited in Claire Selltiz, Marie Jahoda, Morton Deutsch, and Stuart W. Cook, *Research Methods in Social Relations*, New York: Holt, Rinehart, & Winston, 1959, p. 323.

[3] Donald T. Campbell and Joyce Sween, "Tests of Significance for the Interrupted Time-Series Experiment." Mimeographed Research Report, Department of Psychology, Northwestern University, 1965.

[4] A more expanded analysis of the regression problem is contained in Campbell and Clayton, 1961, *op. cit.* note 1 above.

[5] Campbell, 1963, *op. cit.* note 1 above.

[6] See the time-series from the early British Industrial Research Board factory studies, presented in Campbell, 1963, *op. cit.* Most of these were interpreted as showing effects which statistical analyses fail to confirm. Monte-carlo runs of no effect conditions produce frequent series that on visual grounds look to be proof of effect, in the studies reported by Campbell and Sween, *op. cit.* note 3 above.

[7] The F is from Helen M. Walker and J. Lev, *Statistical Inference*, New York: Holt, 1953, pp. 390-395. All tests are presented in more detail in Campbell and Sween, 1965, *op. cit.* note 3 above.

[8] Campbell and Sween, 1965, *op. cit.* note 3 above.

[9] A. M. Mood, *Introduction to the Theory of Statistics*, New York: McGraw Hill, 1950, pp. 297-298.

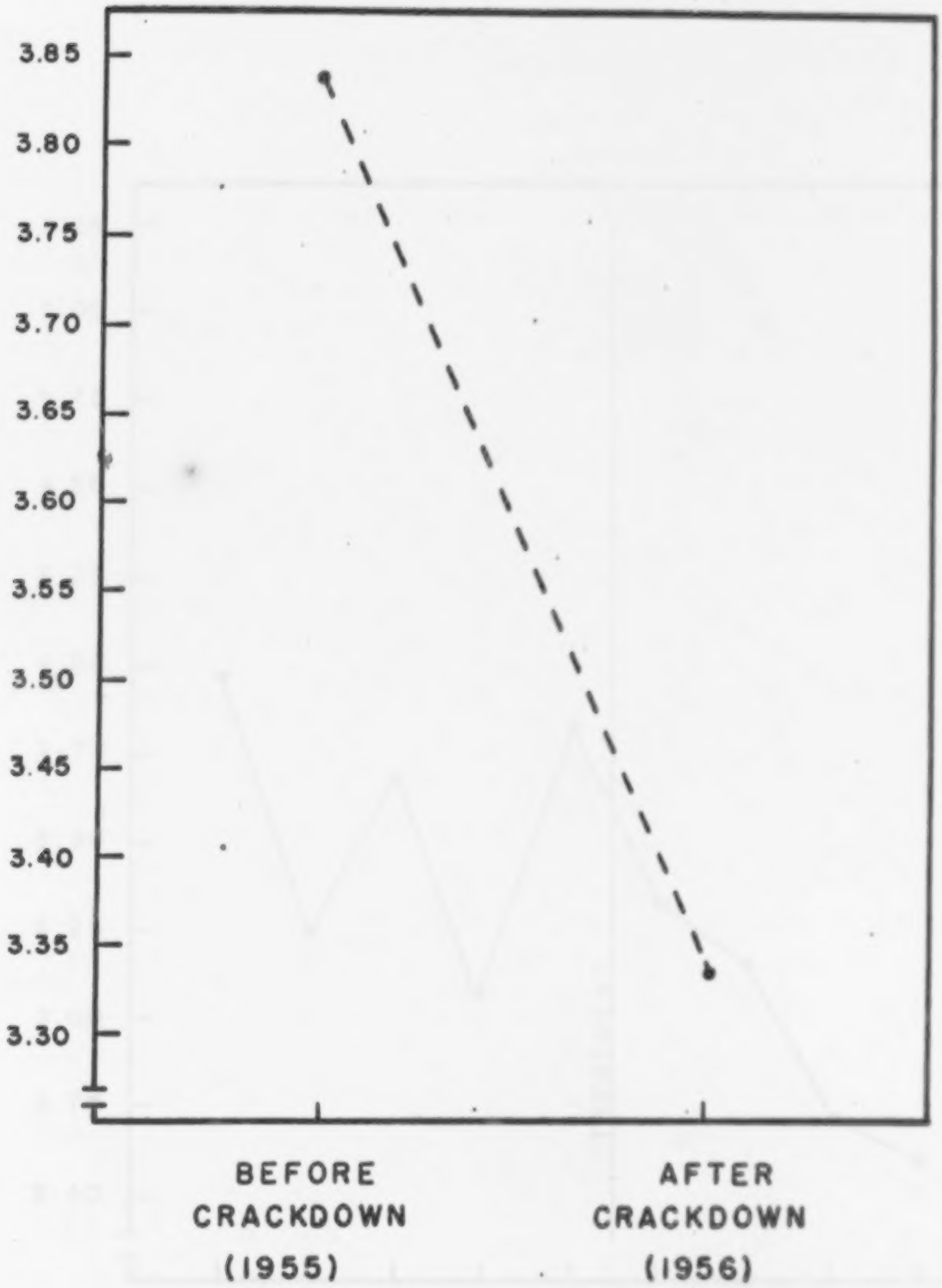


Figure 1. Pre- and Posttest Measures of Connecticut Traffic Fatalities per 100,000,000 Driver Miles

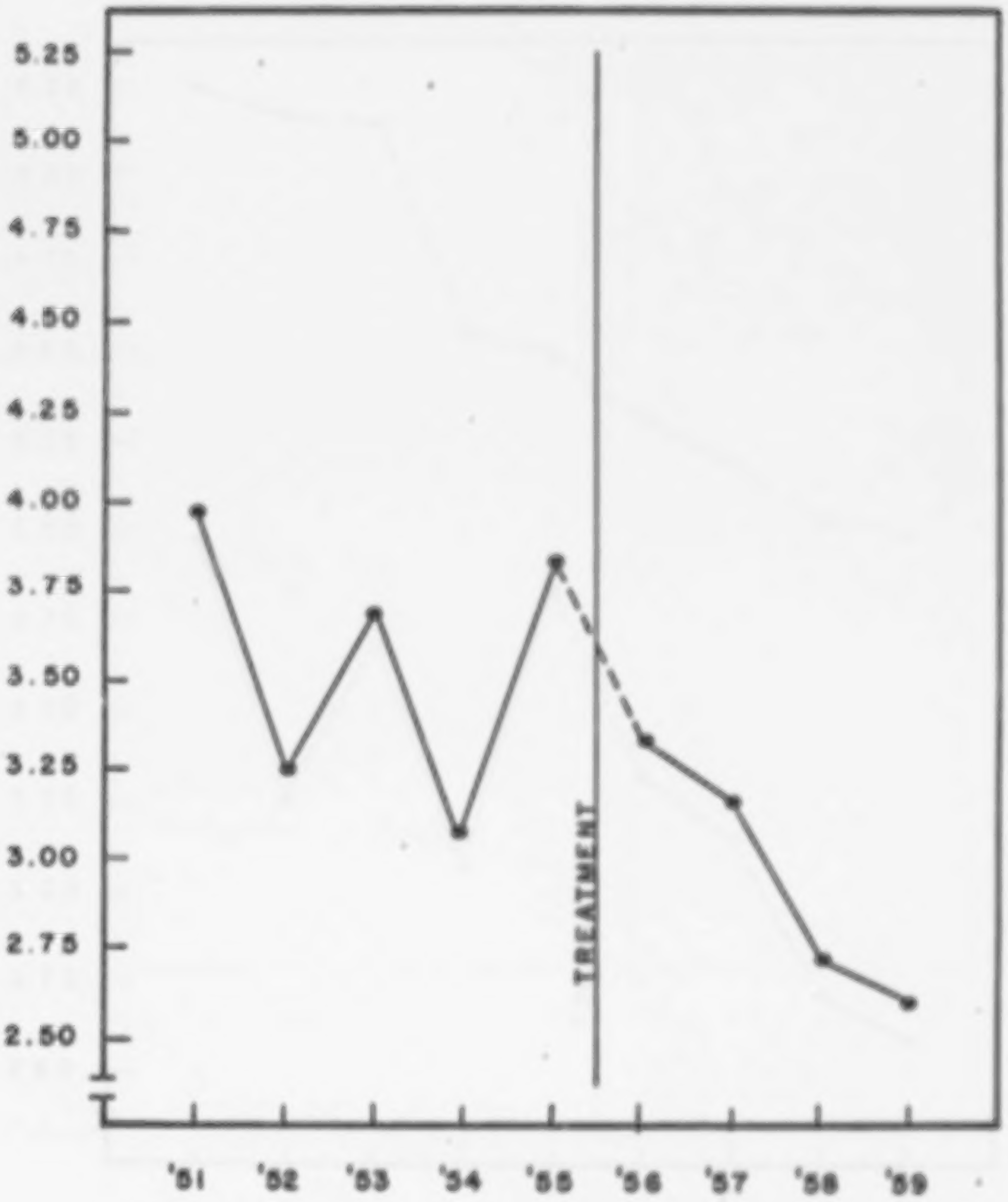


Figure 2. Interrupted Time-Series Presentation of Connecticut Traffic Fatalities

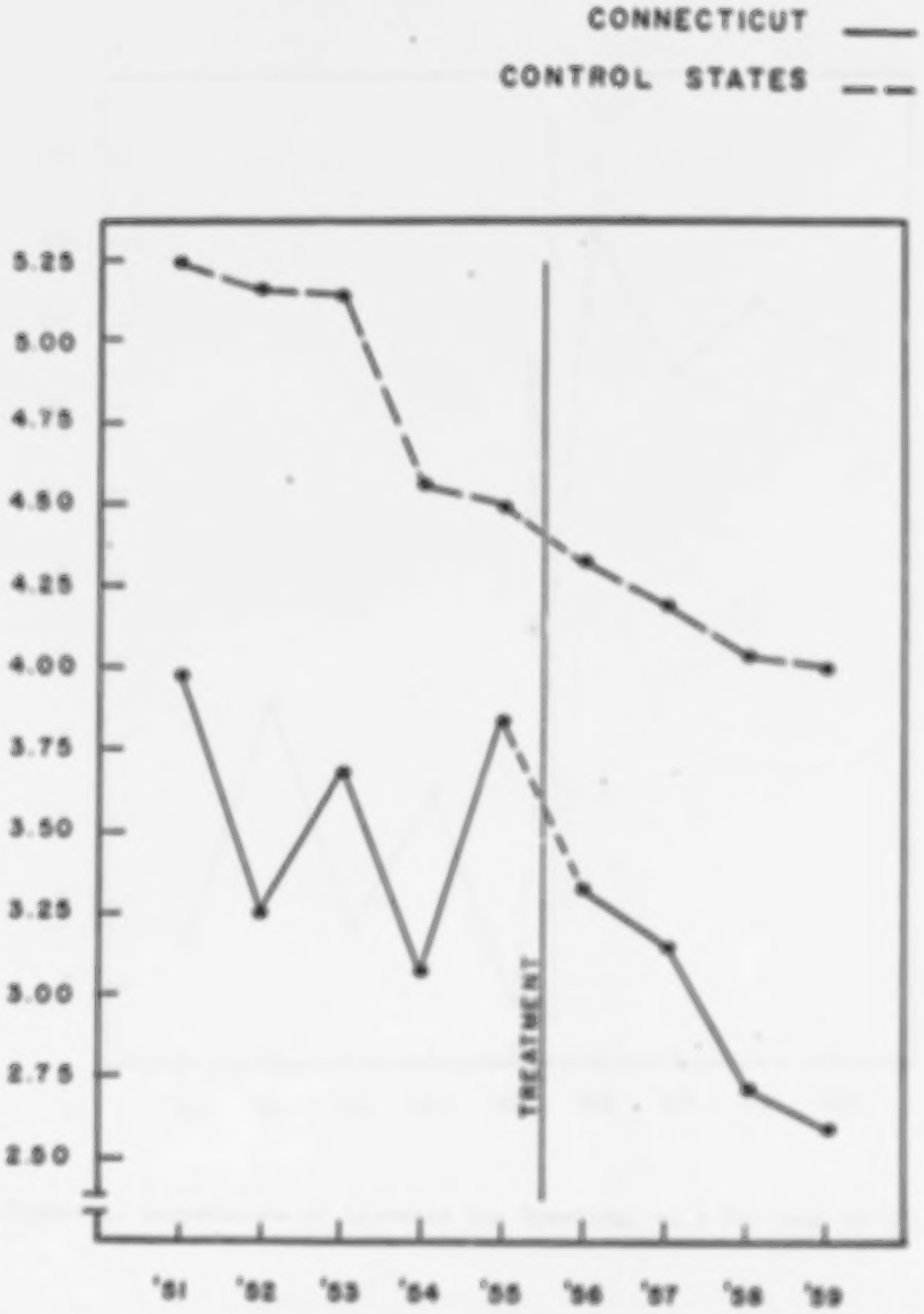


Figure 3. Multiple Time-Series Comparing Connecticut Fatalities with those of Four Comparable States

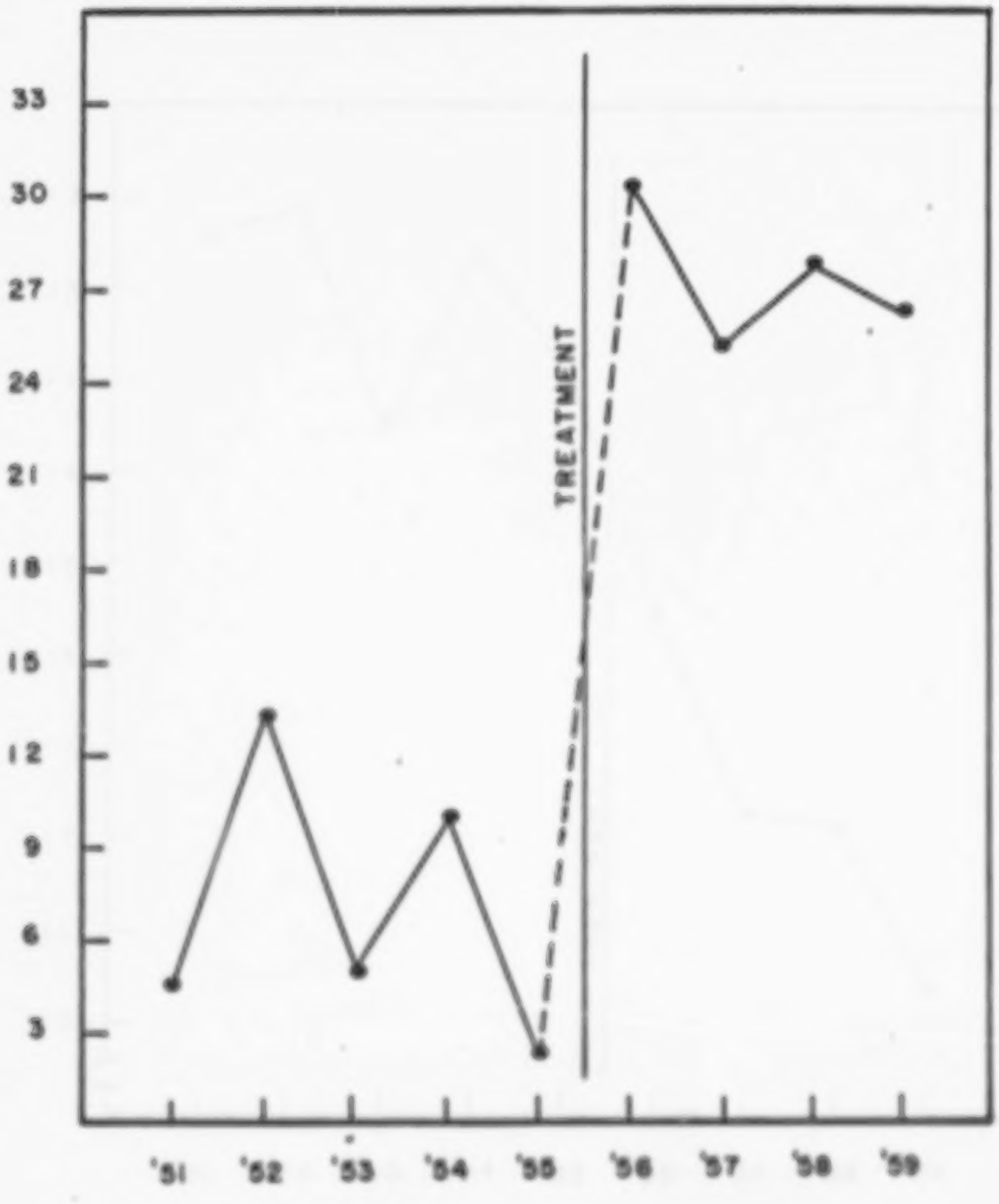


Figure 4. Suspensions of Licenses for Speeding, as a Per Cent of All Suspensions

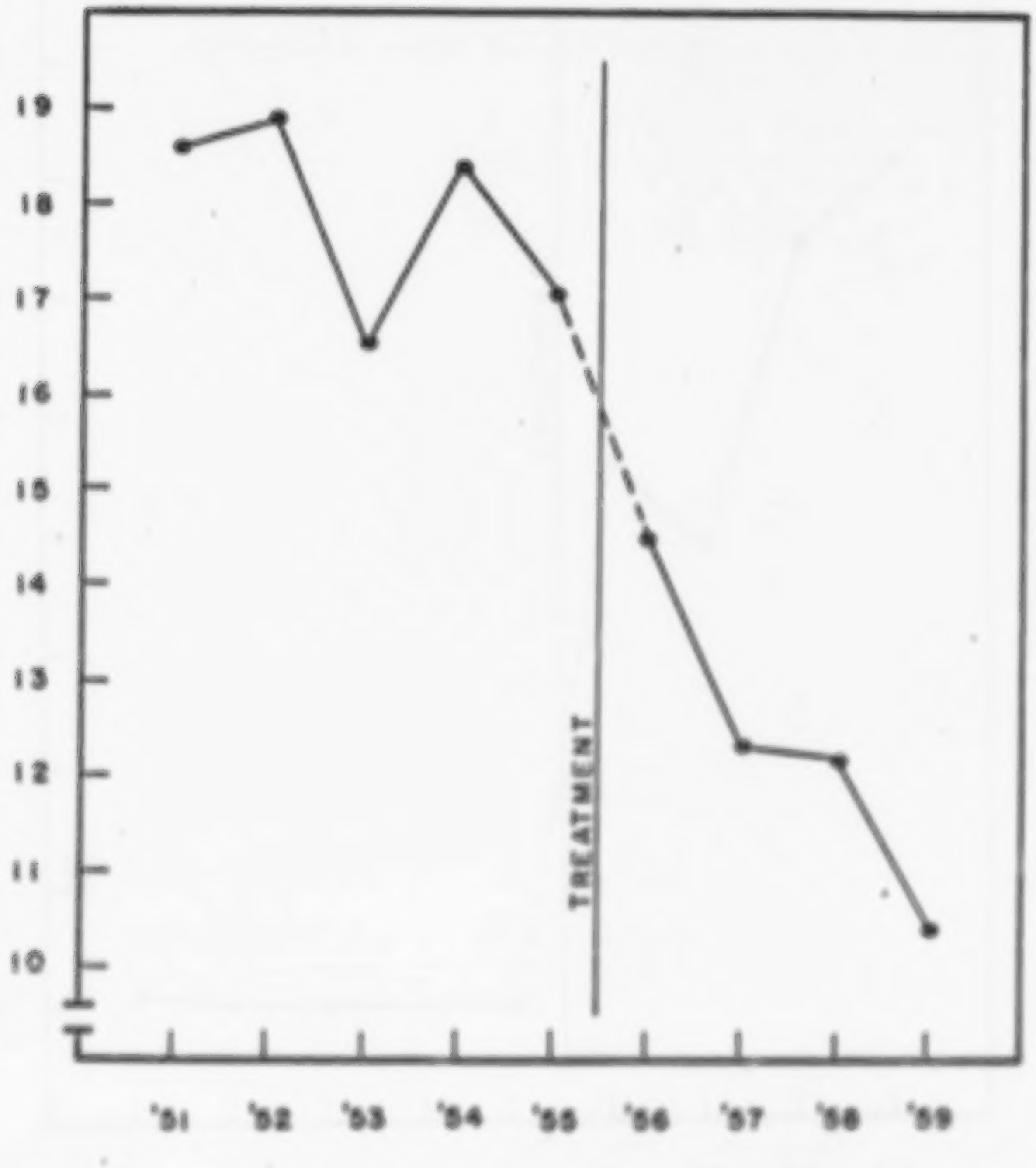


Figure 5. Speeding Violations, as Per Cent of all Traffic Violations

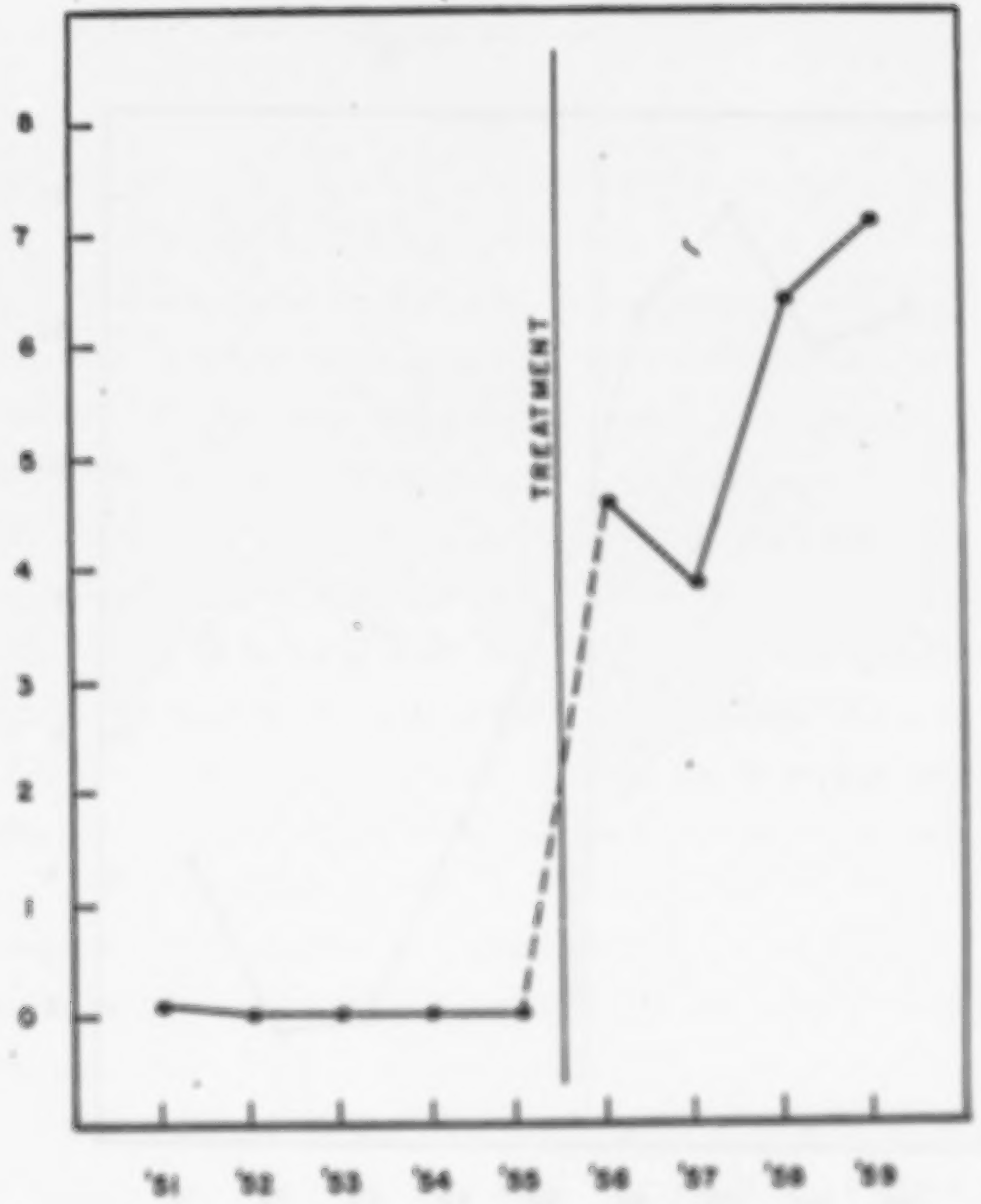


Figure 6. Arrested while Driving with a Suspended License, as Per Cent of Suspensions

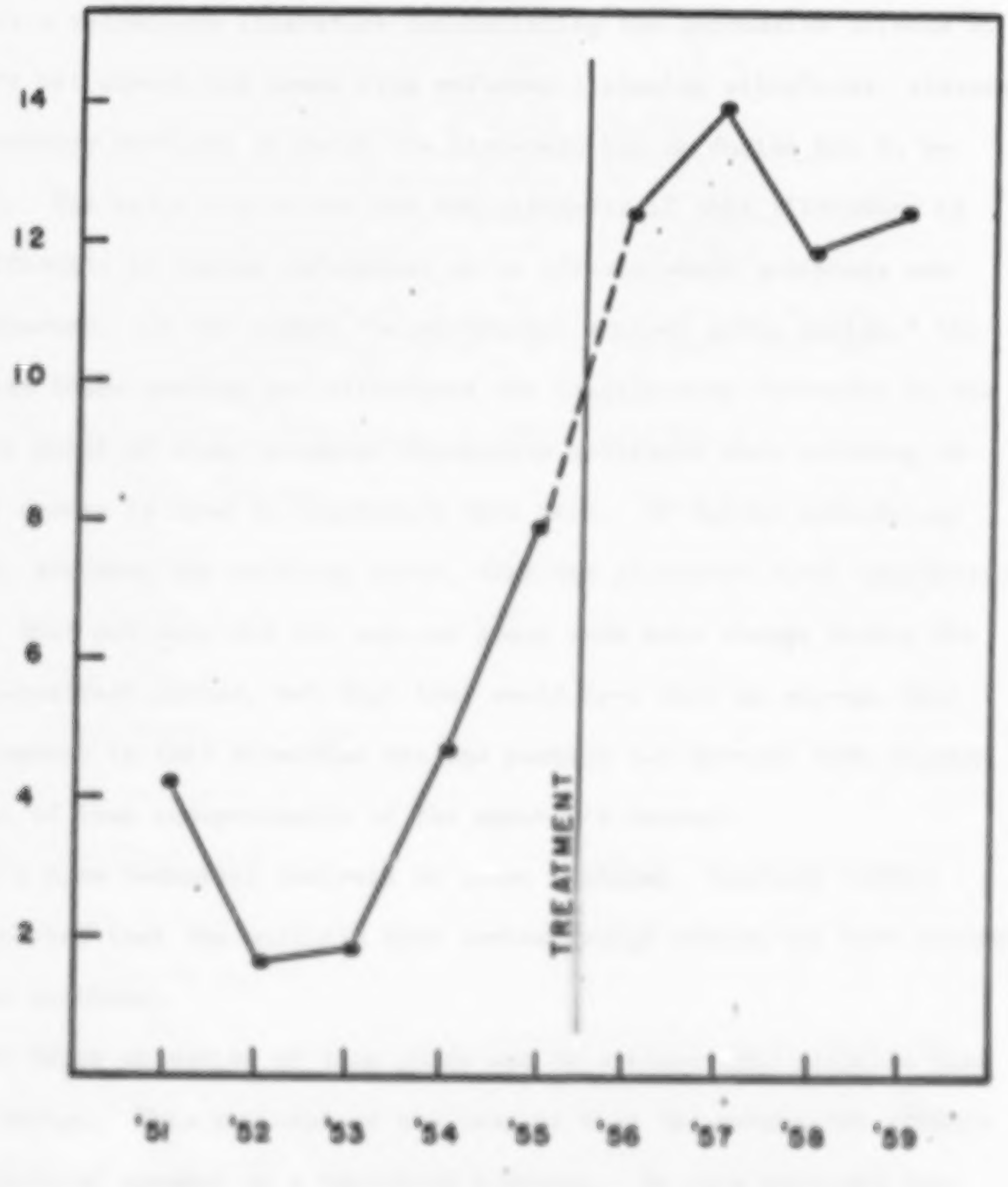


Figure 7. Per Cent of Speeding Violations Judged Not Guilty

A Multiwave Panel Analysis of the Effects of a Political Speaker upon a Self-Selected Audience*

Thomas Mader & Carole R. Siegman
Amherst College Northwestern University

There is a voluminous literature demonstrating the persuasive effects of speakers but almost all comes from enforced listening situations, classrooms or laboratory settings in which the listeners had no choice but to be present. The major reason for the one-sidedness of this literature is the difficulty in making inferences as to effects where audiences are self selected. In the common "nonequivalent control group design," the fact that those seeking out attendance are usually more favorable to the speakers point of view, produces regression artifacts when matching on pretest scores is used to counteract this bias. If better methodology is used, avoiding the matching error, then the plausible rival hypothesis remains that not only did the exposed group show more change during the pretest-posttest period, but that they would have done so anyway--that a development in this direction was and perhaps had already been characteristic of them independently of the speaker's message.

In a more technical analysis of these problems, Campbell (1963), has suggested that the multiple time series design offers the best control of these problems.

The major objective of this study was to explore this multiple time series design. This methodology was used to test the persuasive effects of a political speaker on a valuntary audience. We have employed this

---

*Supported in part by Project C998, Contract 3-20-001, Educational Media Branch, Office of Education, U. S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act, Donald T. Campbell, Northwestern University principal investigator.

design in a "panel" setting by repeating a questionnaire to the same individuals before and after a treatment (the speaker) was applied.

Representationally the multiple time series design appears as the following.

$$\begin{array}{ccccccc} O & O & O & X & O & O & O \\ \hline O & O & O & & O & O & O \end{array}$$

> "*X* will represent the exposure of a group to the experimental variable or event, the effects of which are to be measured; *O* will refer to some process of observation or measurement; the *X*'s and *O*'s in a given row are applied to the same specific persons. The left-to-right dimension indicates the temporal order, the *X*'s and *O*'s vertical to one another being simultaneous..." The parallel rows seperated by dashes indicate *comparison groups not indicated by random assignment*. (Campbell: 1963, pp. 6 & 7)

As can be seen, this design provides two demonstrations of the influence of the stimulus. First, it allows for examination of the stimulus in relation to the pre-*X* values. And secondly, the design provides a check on validity by comparing the exposed group with a comparison group--a group not exposed to the stimulus.

In determining the causal agent in any change situation, one is provided with a more valid indicator as to the nature of the agent, if the prechange conditions have been systematically examined. By doing so, we can illicit any appreciable fluctuations which may have naturally occured and which might be regarded as plausible alternative hypotheses. Thus, the experimental effect becomes clear when seen by examining the conditions before it was applied, as well as the attitudes of the comparison group where there was no treatment.

The implementation of this particular methodological design provided us with means by which we studied the effect of an address which was to be given by William F. Buckley, Jr., at a metropolitan university. Buckley is editor of the Conservative National Review magazine and a reputedly influential speaker. The effect of the speaker on the attitudes of the respondents was measured by the use of a series of attitude questionnaires and was measured by attitude changes of those exposed to him and those not exposed.

METHOD

Subjects: One thousand students were adminstered questionnaires; however, the amount of useful data reduced this amount to one hundred and ninety four. The questionnaires were administered in various classes at St. Johns University in New York; freshmen, sophomores, juniors, and seniors were included as well as students of both sexes.

Materials: The students were given a 16 item attitude test; each item was represented by a pair of opposite, statements. The statements were marked A & B, and the respondent was asked to indicate his attitude toward each item by expressing his choice on a seven-point scale.

For example,

ITEM 1

(A) The United States should cease giving aid to Communist Yugoslavia and Poland.
(B) United States aid to Communist Yugoslavia and Poland is essential at the present time.

A! A a ? b B B!

A!...statement A is strongly preferred over statement B.
A....statement A is preferred over statement B.
a....statement A is somewhat preferred over statement B.
?....I cannot decide between A and B.
b....statement B is somewhat preferred over statement A.
B....statement B is preferred over statement A.
B!...statement B is strongly preferred over statement A.

The alternatives were either liberal or conservative in nature and represented contemporary affairs. They included such items as race and foreign relations, the role of the federal government, the United Nations, and membership and behavior in various organizagions.

Also on the first page or instruction sheet the student was asked for his I.D. number, the means by which questionnaires were identified in successive administrations; year in school, age, and sex was also obtained.* (See Appendix A for copy of questionnaire)

Design: The questionnaires were given in four waves approximately two weeks apart. William Buckley spoke 10 days after the third wave had been administered, and the fourth wave was given during the following four days. Each wave took approximately five days to complete.

Those who completed four waves numbered seventy-one; seventeen of these heard Buckley speak. Those completing only two waves (three and four--the pre and post X) totalled 88; eighteen heard Buckley. Other combinations have not been analyzed.

Buckley spoke for a little over one half hour on "The Real Meaning of the American Right Wing"--primarily an attack on Norman Mailer. There was a one hour and fifteen minute question period following. He covered seven out of the sixteen items on the questionnaire.

Procedure: Students in regular classroom meetings were given the questionnaire booklet which included self-administering instructions and information that they would be receiving a similar instrument every two weeks during the fall semester.

---

*Also, included were two additional pages; the first was a Survey of Reading and Listening Habits which included a list of 94 magazines, newspapers, magazine and newspaper columnists and television programs ranging from Conservative to Liberal. The second page was a Shift of Opinion-Source of Influence questionnaire. The students were asked,

> Do you think that any magazine, newspaper, columnist, television program, public lecturer, professor, governmental decision, etc., has influenced you within the past *two* weeks--in CHANGING *or* RE-ENFORCING your opinion on some item or items covered in the questionnaire?

The respondent was asked if he was a member of a fraternity or sorority and his political party preference.

These data will be analyzed in the Ph.D. dissertation in preparation of Thomas Mader. It is tentatively titled, A Descriptive Analysis of the Speaking of William F. Buckley, Jr.

Data Analysis & Discussion of Results

In order to determine Buckley's influence upon the students' attitudes the data was seperated into two groups, those who attended the lecture and those who did not; the latter group providing the comparison group. Mean scores for each item of the four wave and two wave questionnaires were calculated and graphically plotted. (See Appendix B) This made it possible to observe; (1) what the pre-Buckley political attitudes and differences were of the students who did attend compared with those who did not attend, (2) the influence of Buckley upon those attending his address vs. the comparison group indicated by the change, and (3) the general trend of attitudes of both groups over time.

Since Buckley was judged to be an articulate spokesman for conservatism, it became clear that the most pertinent manner to judge attitude change would be to use precisely those test items which would indicate liberal and conservative attitudes. This would allow for a clearer picture not only of change due to the stimulus but of the pre-existing attitudes of the subjects.

Pre-Buckley Attitudes

In terms of pre-existing attitudes, one of the more important findings is the extent to which the student body, as represented in the sample, is conservative in political attitude. In the four and two wave series of questionnaires we find that ten out of the sixteen items (62%) were answered on the conservative side of the scale. This includes both those who did and did not attend Buckley's lecture.

For the four wave people the mean score for all the pre-X items answered on the conservative end of the continuum was 2.70. A seven point scale was used, one being at the conservative end of the continuum and seven at the liberal end. This indicates a slight more liberality on the part of those who did not attend. The mean values for the two wave people were 2.85 for those who attended the lecture and for those who did not it was 3.03. Since these are average scores across all items it should be realized that when items are viewed individually a greater distance between the two groups may be seen. See graph 3 for example.

Insert Graph 3 about here

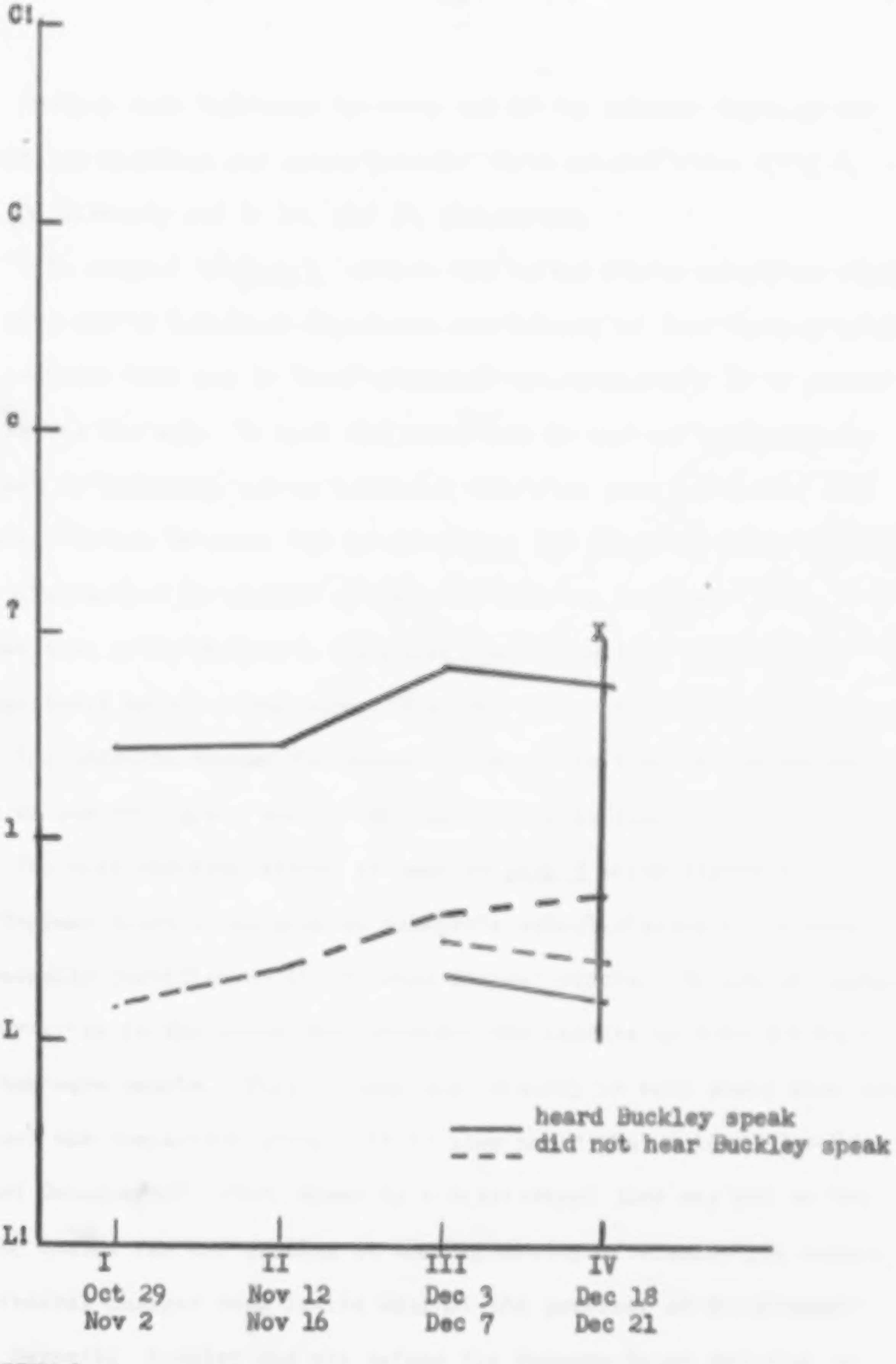
Items 2, 3, 5, and 9 were rated liberal by both the four and two wave people, as well as by both those who attended and did not. For item 6 on the four wave group both groups were found around the neutral point. And on item 13 on the two wave questionnaire, we find those who attended were on the conservative side of the continuum and those who did not were on the liberal side.

Thus, it becomes apparent that in general the pre-X attitudes of the subjects were similar to that of the speaker, for both those who did and did not attend.

Buckley's address and chronological events

In our effort to examine the effectiveness of Buckley's lecture we will turn to the address itself and also discuss some historical events and their relevance for attitude change.

A chronology of relevant events was recorded between October 29 and December 21, this being the time period that the four waves of questionnaires were administered. The Cuban blockade and James Meredith's entrance into the University of Mississippi were the most profound events during this time.



ITEM 3

L (A) The United States would be unwise to withdraw diplomatic recognition from Communist countries.

C (B) The United States should withdraw diplomatic recognition from all Communist countries, including Russia.

Buckley made reference to seven out of the sixteen items in his speech and question and answer period. He discussed items 1, 2, 4, and 13, directly and 7, 10, and 14, indirectly.

With respect to Item 1, whether the United States should or should not give aid to Communist Yugoslavia and Poland, we find Buckley taking the position that aid to these countries was permissible if we placed strings on the aid. He made the point that he was not categorically opposed to extending aid to Communist countries that had broken with Russia. As can be seen, the pre-I values, for those who heard Buckley, show fluctuation in opinion as compared with the "control" group. We see that after Buckley's expressed opinion on this subject they change their opinion once more. However, it is inappropriate to conclude that the stimulus caused the change since at the time of the second wave we see this group was at the same 3.29 position.

The most striking effect is seen in Item 2 which stated that the Supreme Court's decision to integrate school classes is legally and morally justified or it violates states' rights. We see an increasing conservatism in the group that attended the lecture in both the four and two wave people. This is seen most clearly in both cases when viewed against the comparison group. It is also noted that between November 16 and December 21, 1962, blame by a Mississippi jury was put on the United States for the rioting at the University of Mississippi campus, and federal charges were levied against the governor of Mississippi, Ross Barnett. Buckley did not defend the Supreme Court decision on the groups that this type of legislation is coercive and does not change individuals' attitudes.

Item 4 deals with the Cuban problem, specifically with the matter of deposing Castro. About half the speech was devoted to the Cuban situation. By this time President Kennedy had undertaken the Cuban blockade (October 22, 1962) and Russians had subsequently removed their missiles from Cuba (October 28, 1962). Buckley said in his speech: "Cuba is a symbol of American Liberalism's failure to meet the challenges of the modern world." He further noted that Kennedy "appears officially to have abandoned the effort to unseat Castro," although Buckley never said that Castro should be unseated. There is a very slight effect on this item for both the two and four waves.

Item 13 states that the United States and Russia should either negotiate a settlement about Berlin or tear down the Wall. With respect to the four wave people who attended the lecture, we see an increasing conservatism in the first three waves and a slight effect or change toward liberality after Buckley's address. The comparison group also had an increasing conservatism but which continued in the fourth wave also. Interestingly, the two wave questionnaires show an opposite effect indicating that a simple pre test post test design excludes the over all picture of the increasing conservatism which in fact changed after Buckley spoke probably indicating that Buckley had no effect. Buckley indicated that those who would negotiate a settlement with Russia were "brave cowards," and that the United States disappointed those people in East Germany, Cuba, Laos, and China who desire freedom. Since Buckley mentioned Items 7, 10, and 14 in general and did not specifically make reference to the problems posed in the questionnaires, these will be discussed later.

Further analysis of items

We will turn now to the other items and in doing so a number of factors emerge which point up some problematical areas in this design. The lack of control over the stimulus was one situation which became manifest since we see that Buckley only spoke on four out of the sixteen items. The results for the remaining twelve are found below.

Insert figures 1-16 about here

For the four wave people

a) *no* effects - Items 3, 7, 11, 14, and 15

b) *slight* effects - Item 5

c) *positive* effects - Items 6, 9, 10, 12, and 16

d) *negative* effects - Item 8

For the two people

a) *no* effects - Items 3, 9, and 12

b) *slight* effects - Item 8 and 11

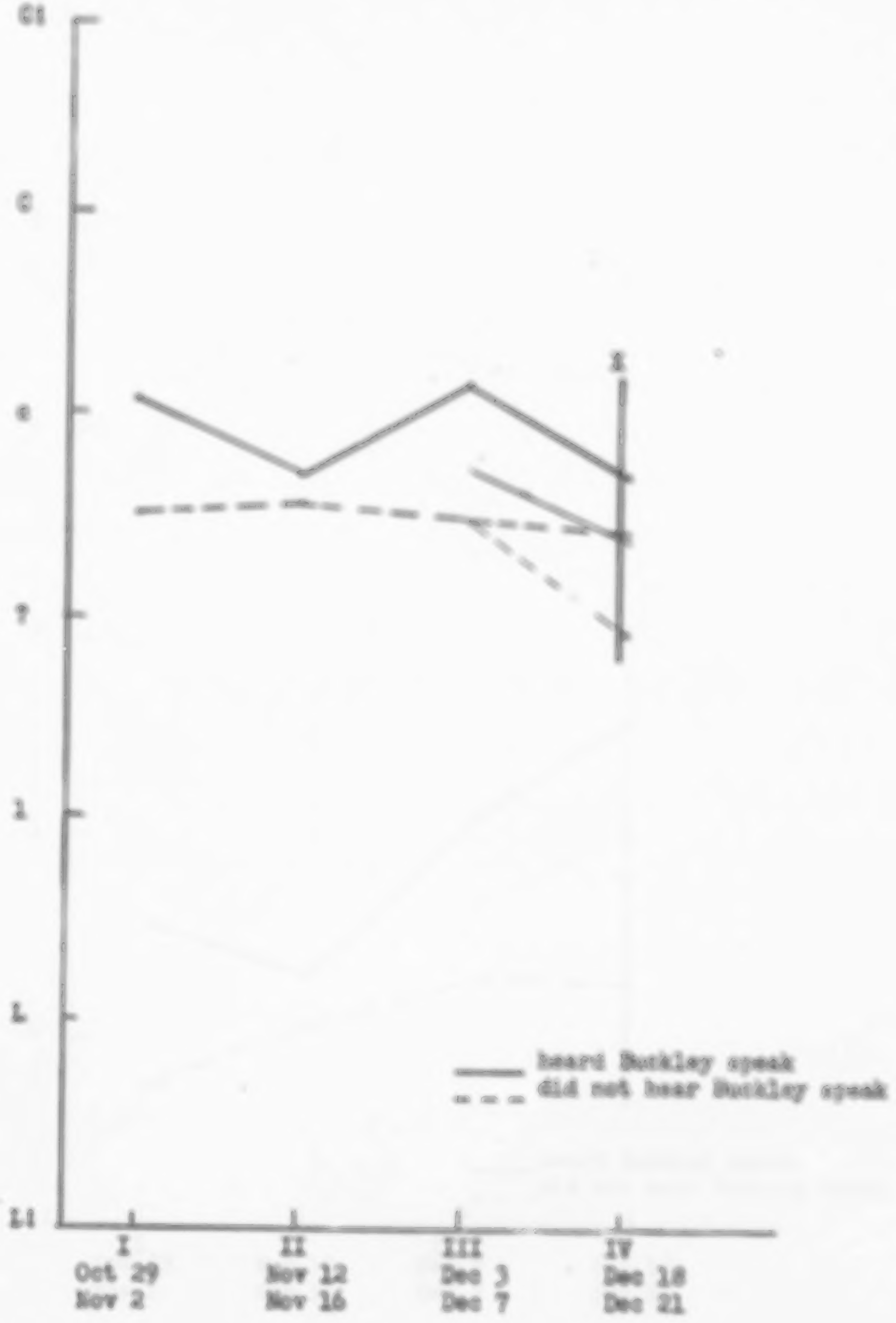
c) *positive* effects - Items 5, 7, and 15

d) *negative* effects - Items 6, 10, 14, and 16

How do we explain these attitude changes or lack of them between the pre and post Buckley scores on items to which he did not refer. To indicate the underlying cause for the diversity of results, the sample itself emerges as an important factor. As was stated earlier, by examining the students' pre-existing attitudes we found the student body to be conservative, and the treatment group slightly more conservative. The interaction between the selective audience and stimulus produced a bias and accounted for the negligible and negative effects.

The use of this design in the voluntary audience setting revealed the need to have more rigid controls over the test-retest situation. This manner of collecting data provided a high degree of "experimental mortality" or a loss of respondents from the initial sample.

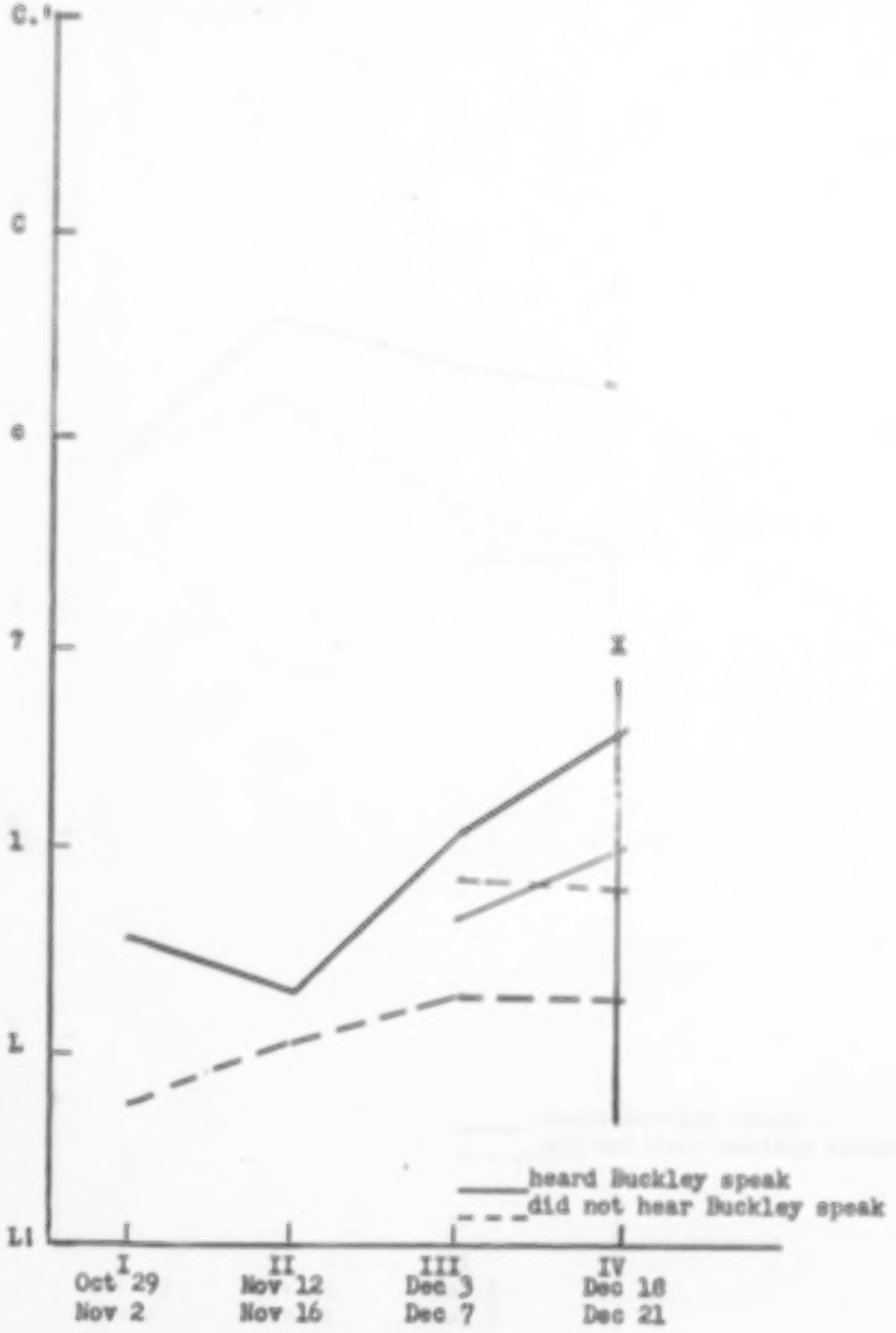
Historical events should be noted for their possible influence upon attitude change. The multiple time series design provides for



ITEM I

C (A) The United States should cease giving aid to Communist Yugoslavia and Poland.

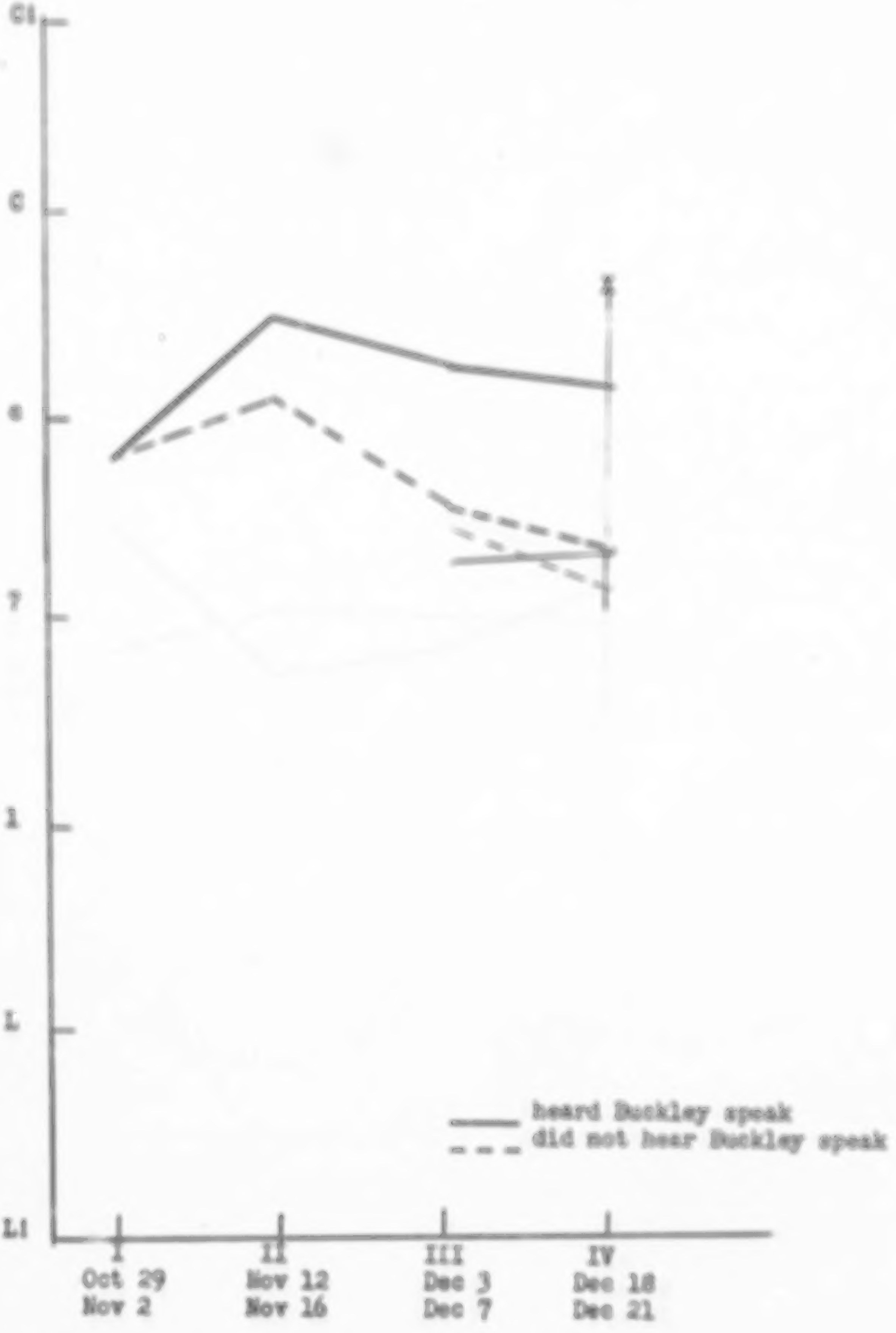
L (B) United States aid to Communist Yugoslavia and Poland is essential at the present time.



ITEM 2

C (A) The Supreme Court violated constitutional provisions for states' rights by compelling schools to integrate their classes.

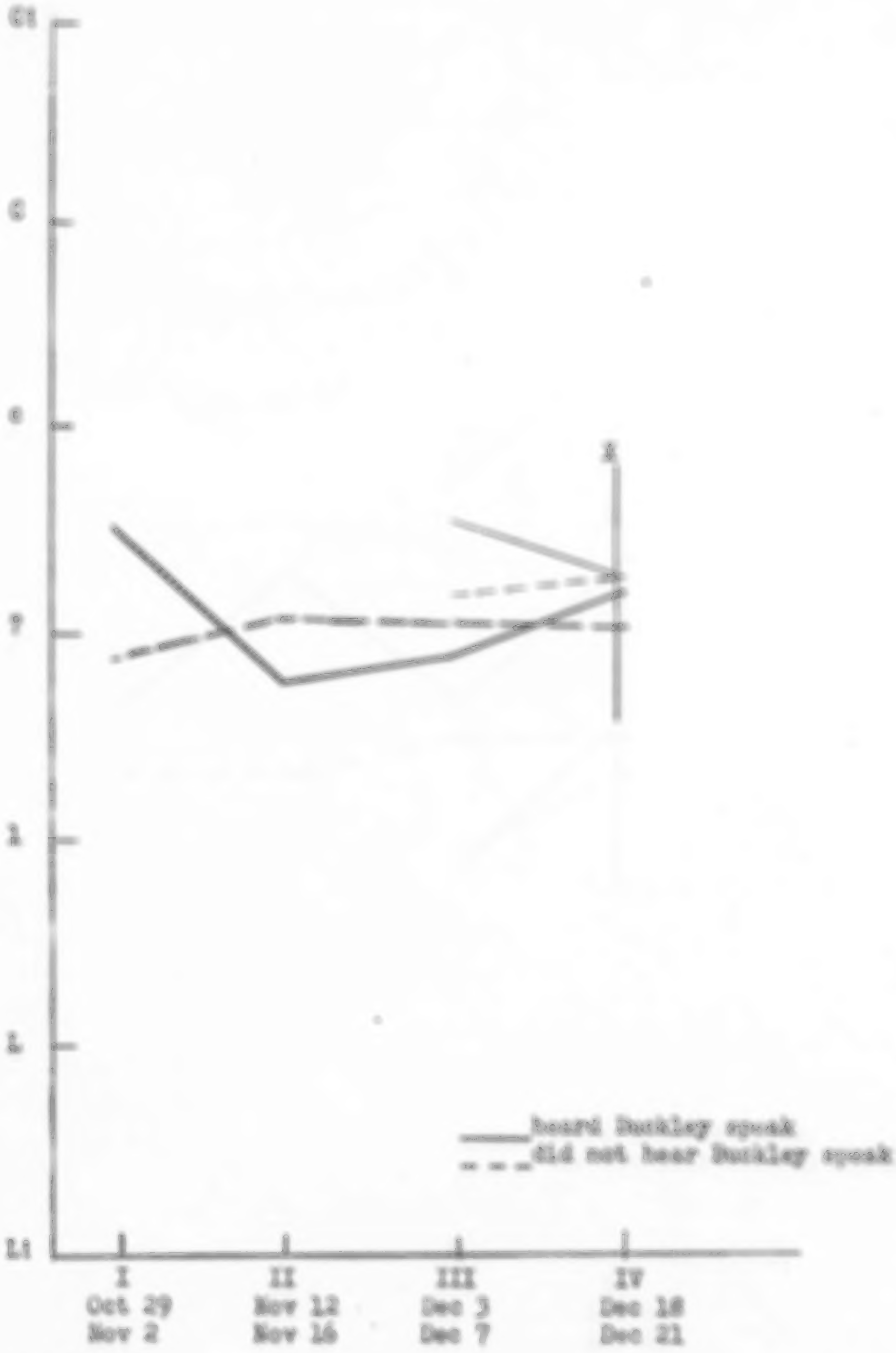
L (B) The Supreme Court decision regarding the integration of public schools was legally and morally justified.



ITEM 4

C (A) The United States should depose Fidel Castro of Cuba.

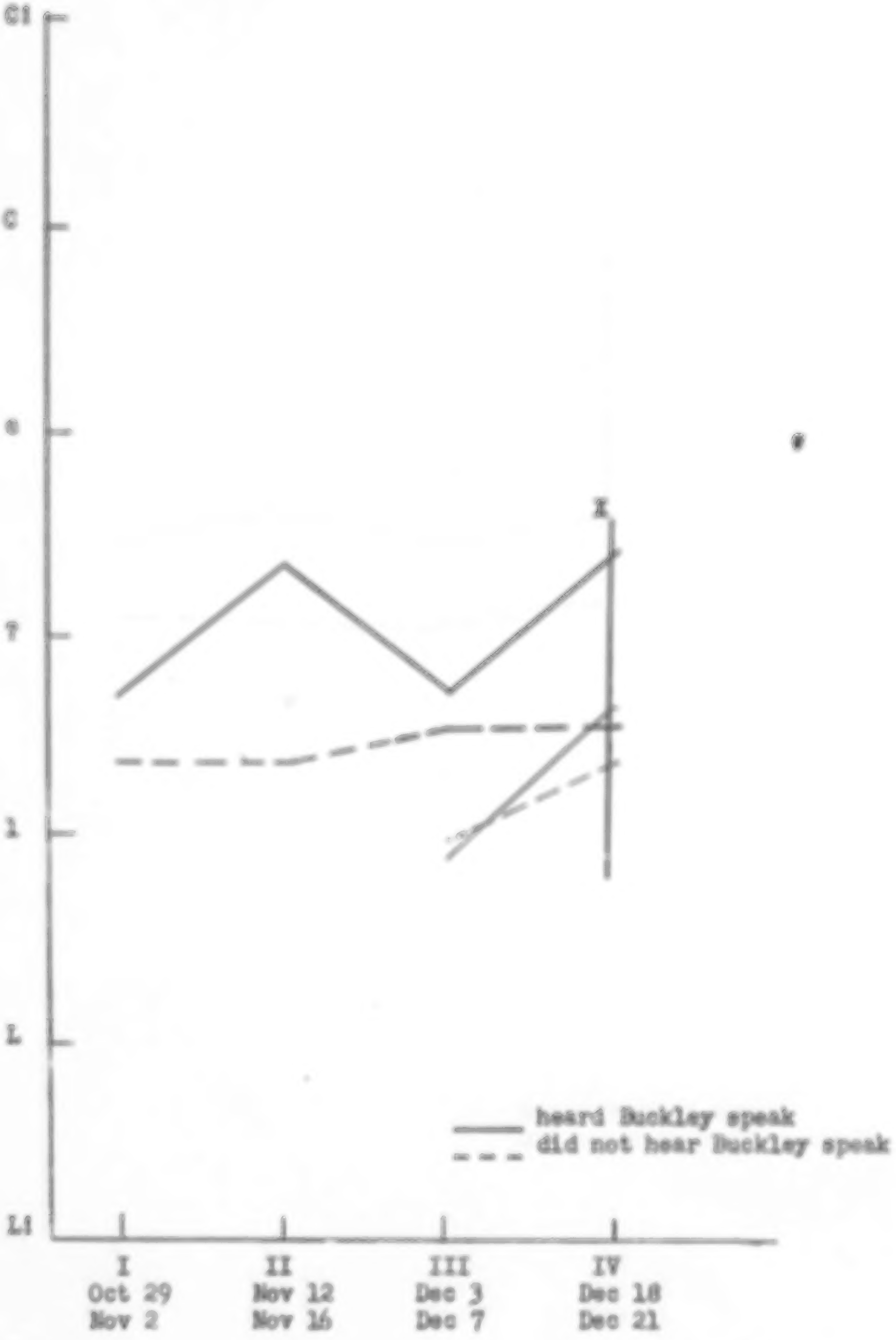
L (B) It would be unwise for the United States to depose Castro.



ITEM 6

L (A) The Federal government is completely justified in placing a 90% tax on incomes over $100,000.

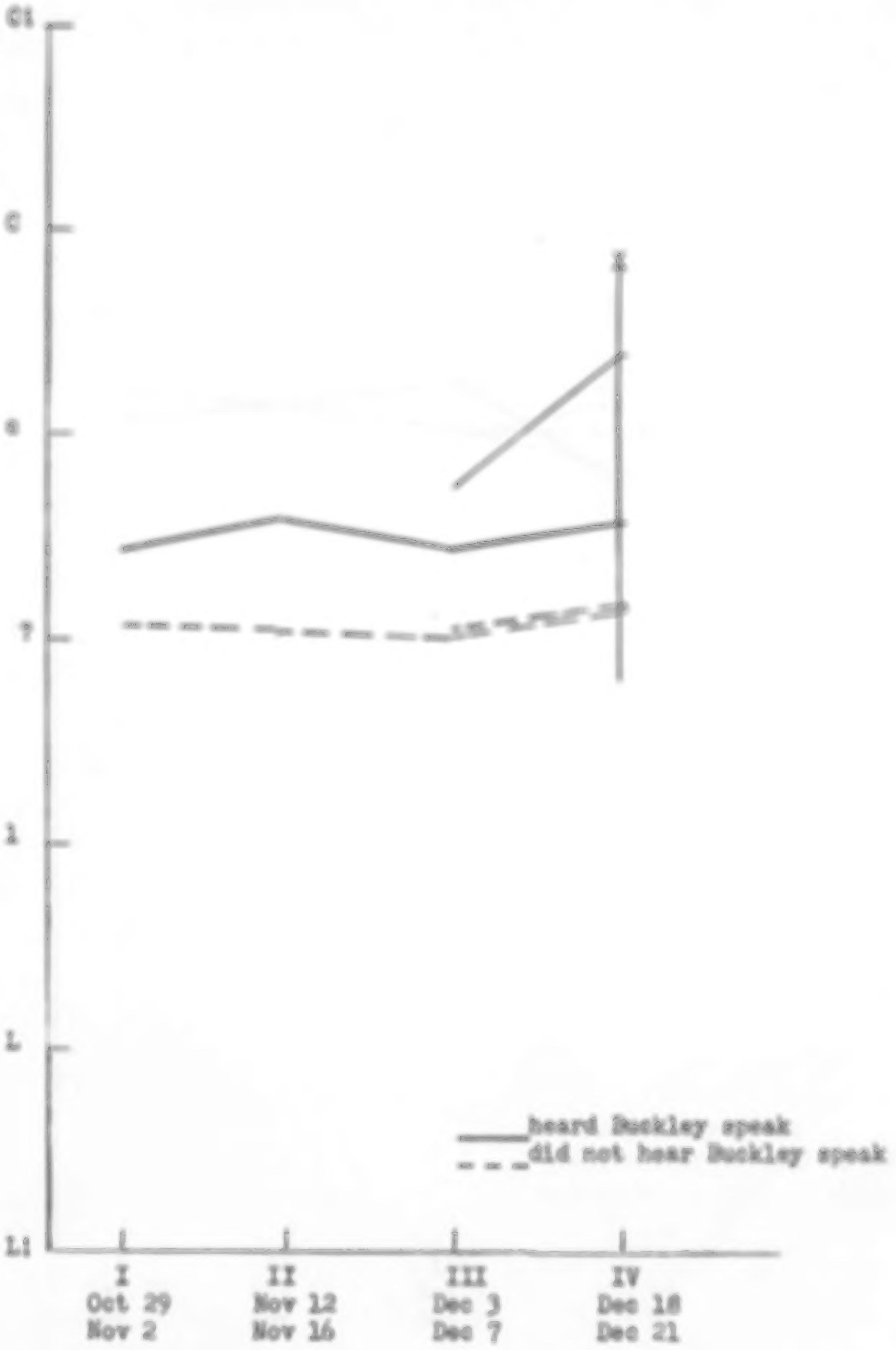
C (B) The Federal government has a right to claim an equal percentage of each man's wealth, and no more.



ITEM 5

L (A) The Federal government should extend financial aid to public schools.

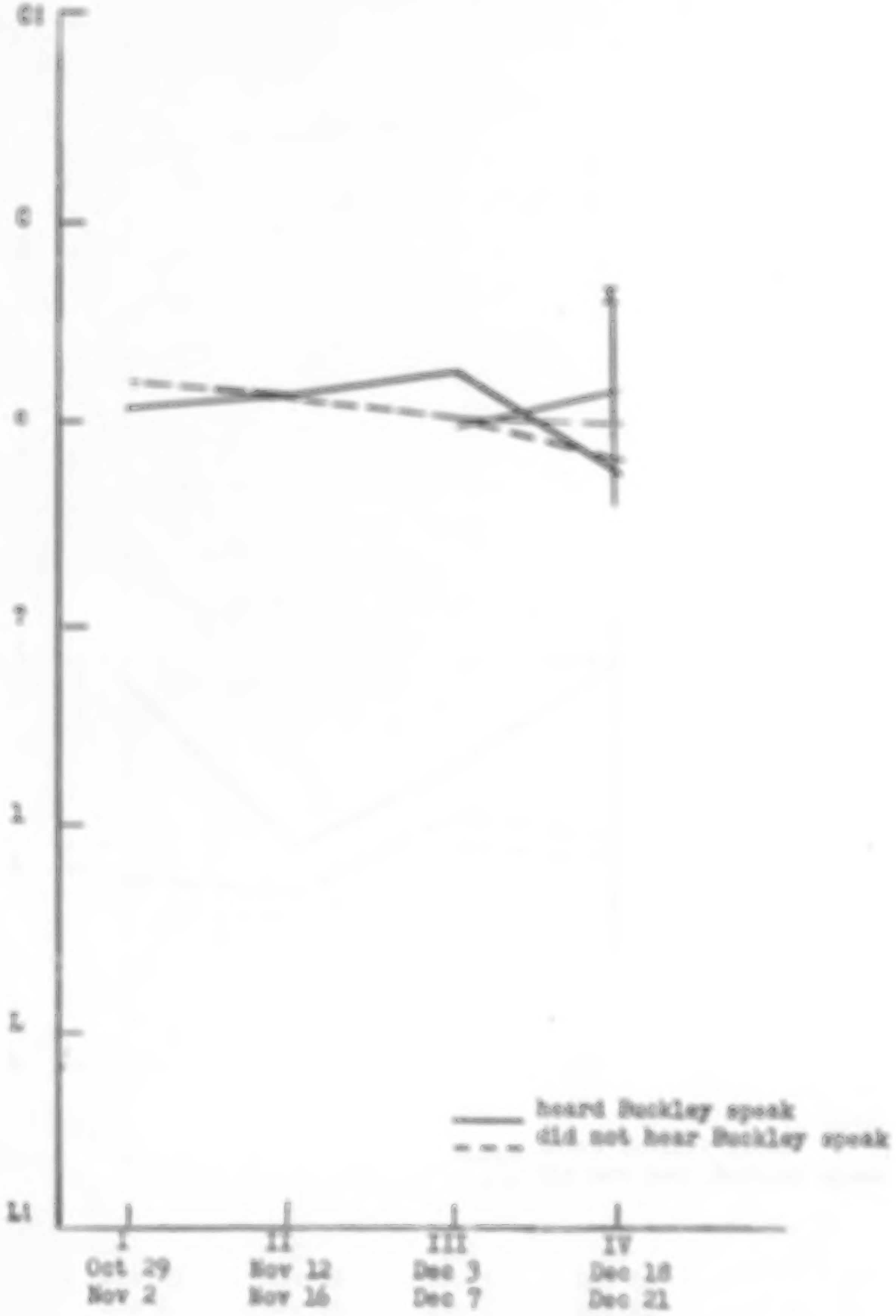
C (B) Each locality should support its own public schools without any Federal assistance.



ITEM 7

C (A) Katanga should not be compelled to submit to the Congolese government.

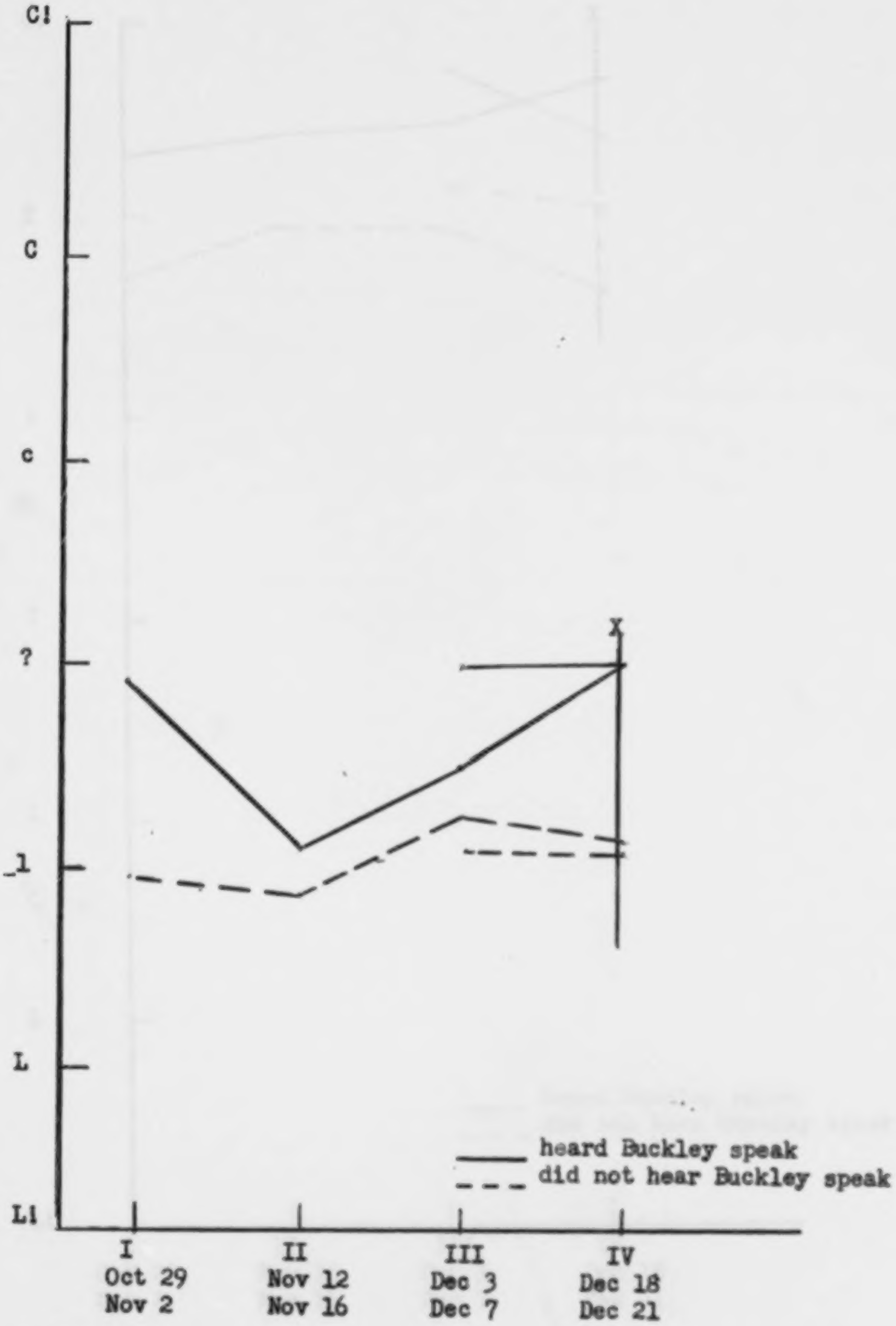
L (B) Katanga is in no way justified in separating itself from the legitimate government of the Congo.



ITEM 8

L (A) The granting of independence to various African nations is desirable.

C (B) A large number of African nations are completely incapable of governing themselves.

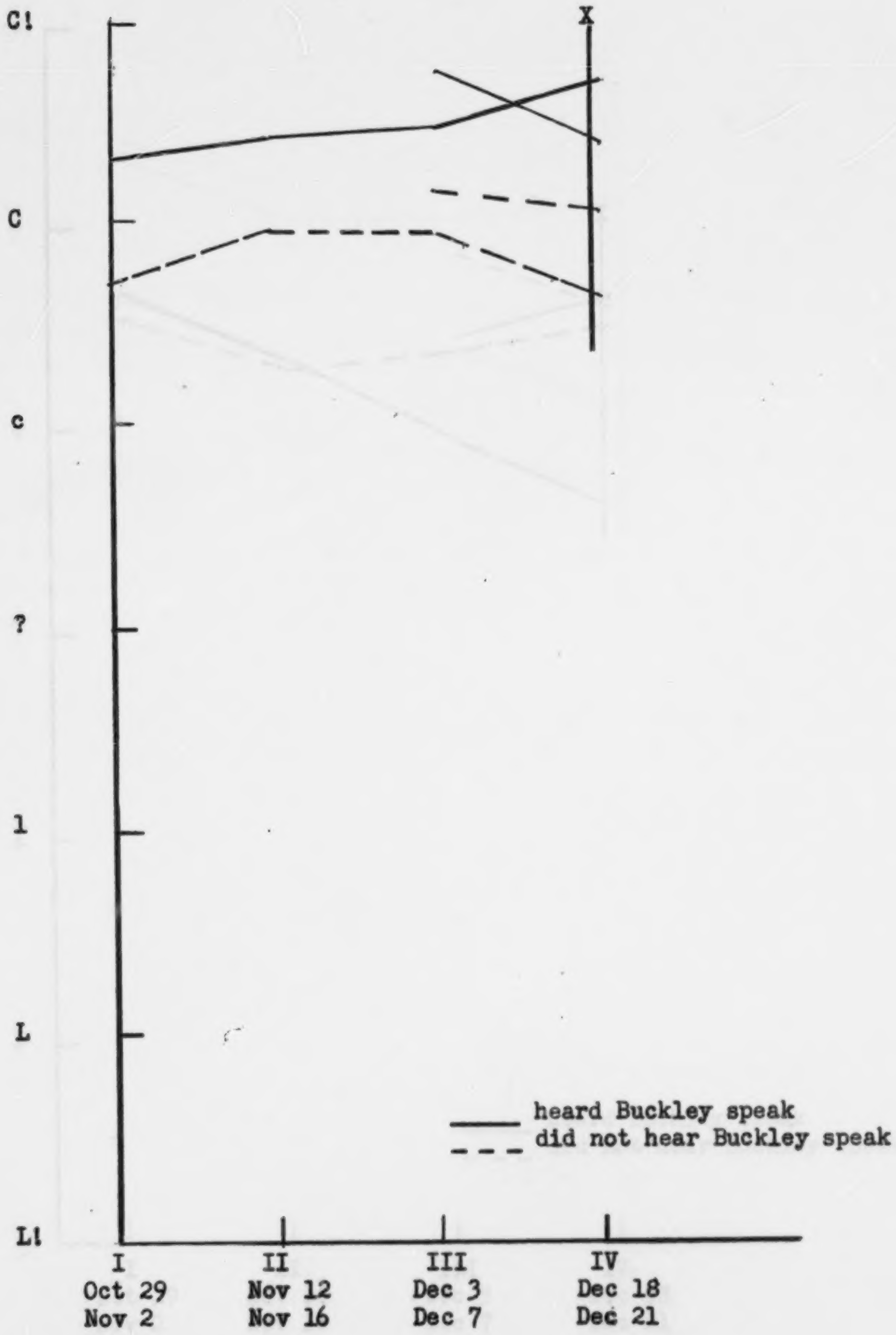


ITEM 9

C (A) Any government program to provide free medical care for the aged should be opposed.

L (B) A government program of free medical care for the aged is obviously necessary.

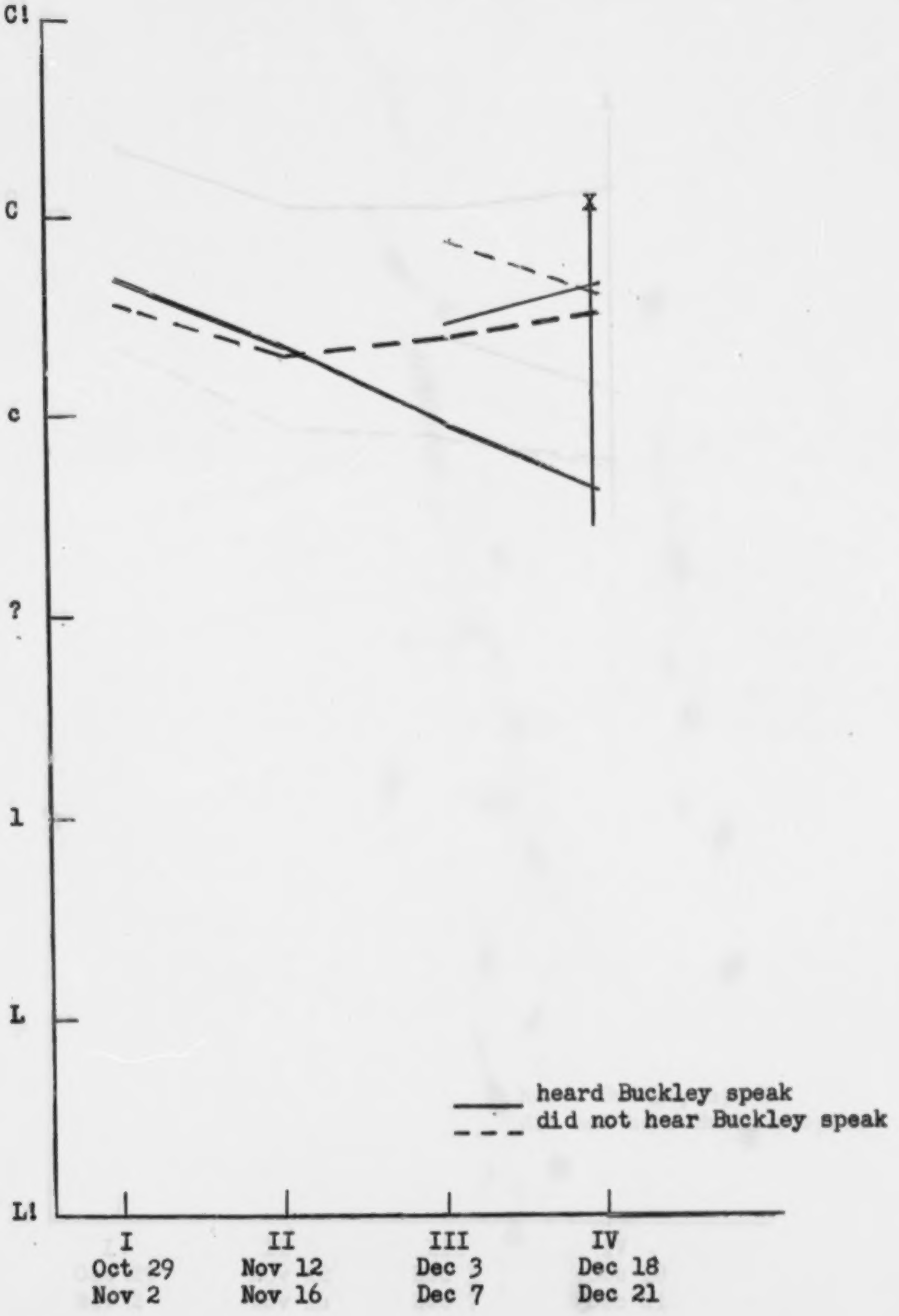
5



ITEM 10

L (A) Red China should be admitted to the United Nations.

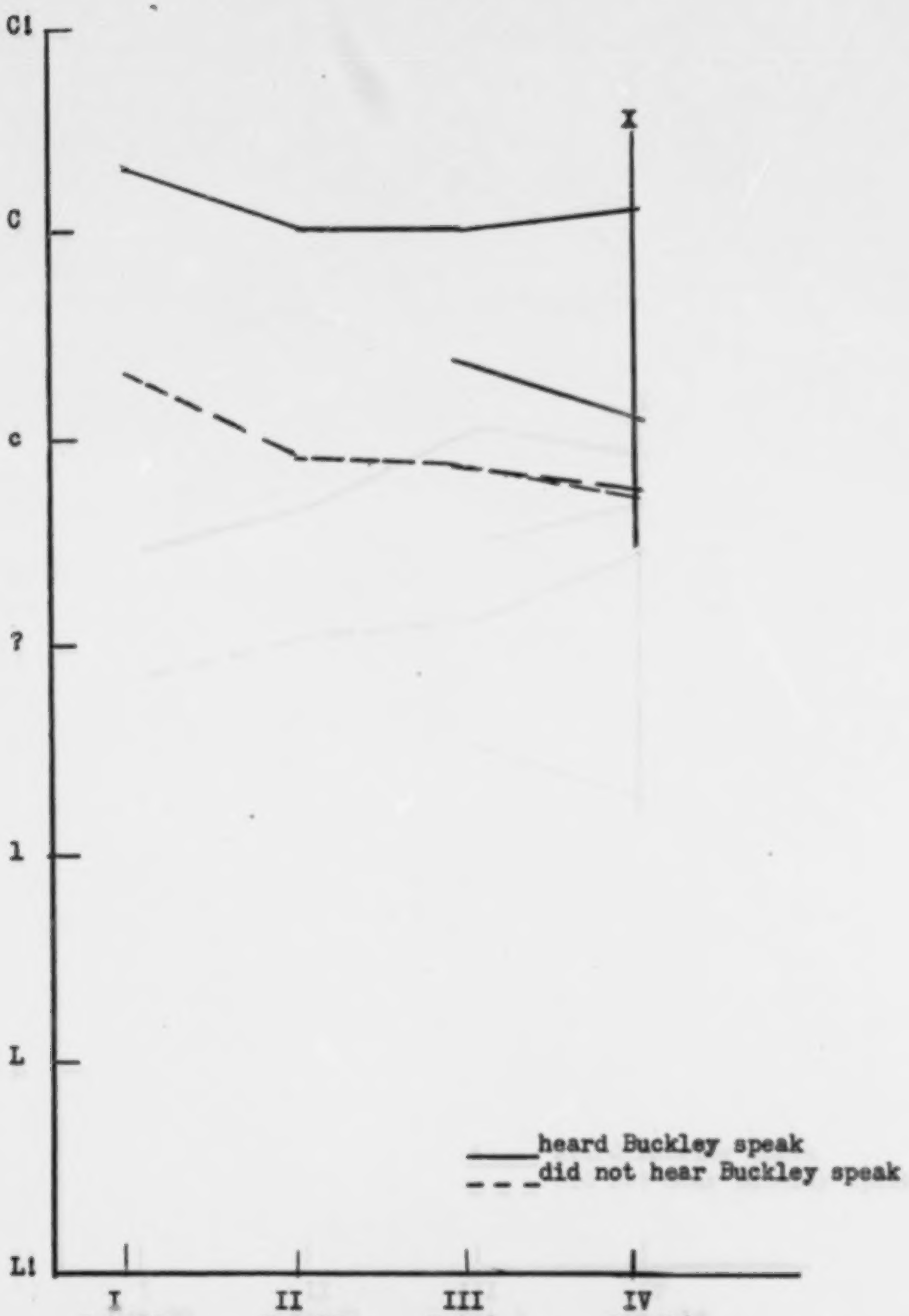
C (B) Red China should NOT be admitted to the United Nations.



ITEM 11

C (A) The Supreme Court decision regarding prayers in public schools distorts the meaning of the United States Constitution.

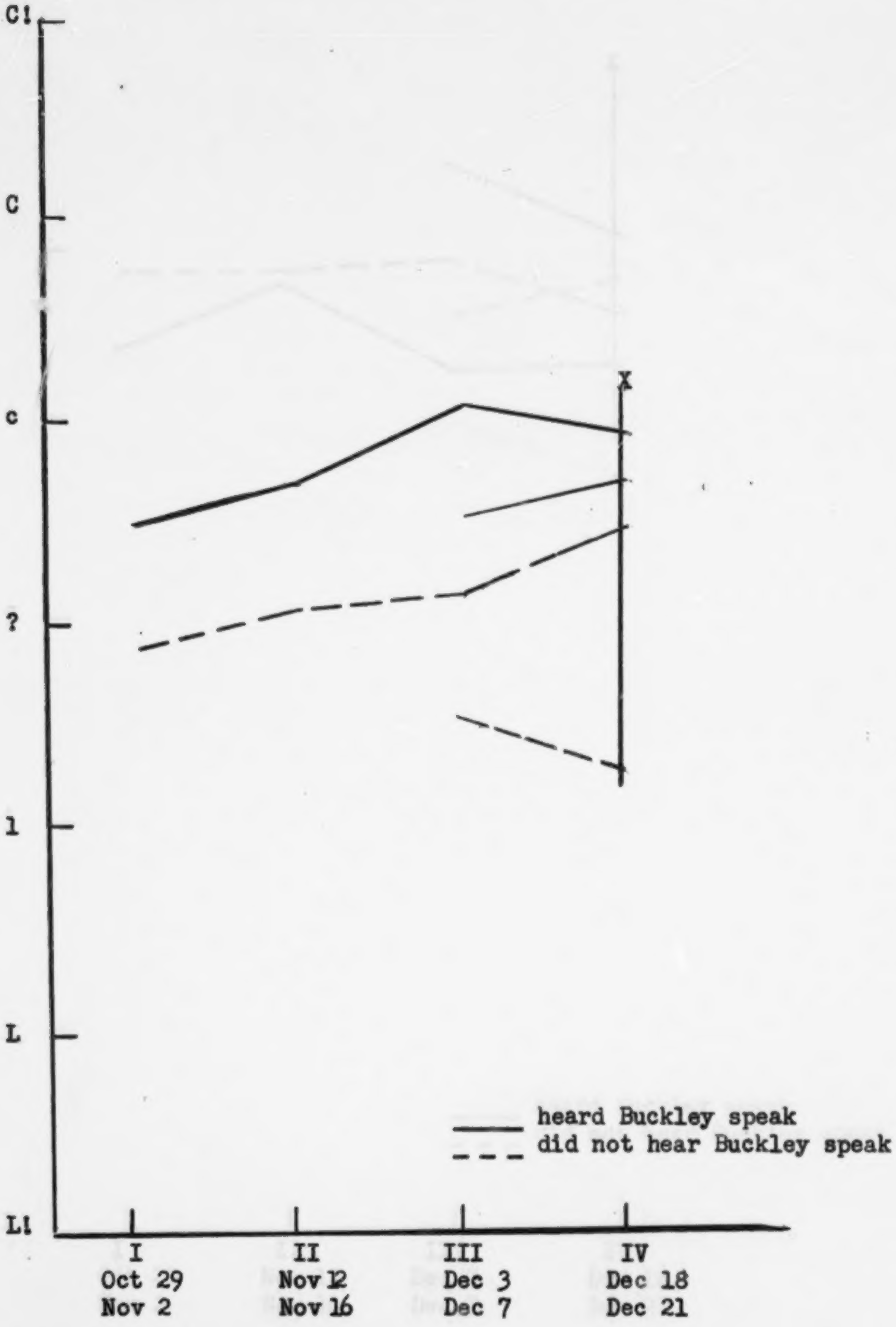
L (B) The Supreme Court decision prohibiting the saying of prayers in public schools was morally and constitutionally justified.



ITEM 12

C (A) The United States should not commit itself to any trade agreement with the Soviet Union.

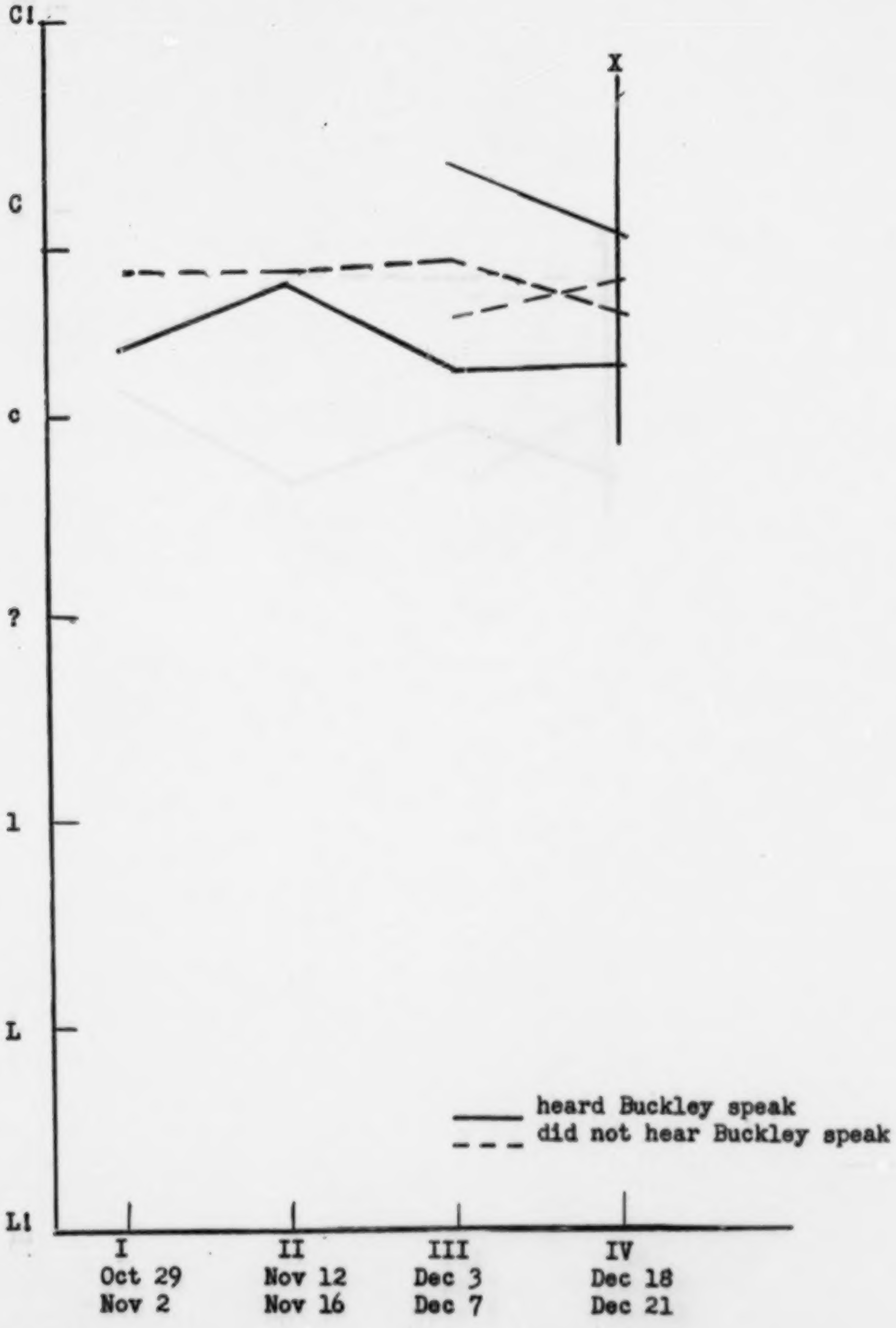
L (B) The United States should liberalize its restrictive trade policy vis-a-vis the Soviet Union.



ITEM 13

L (A) The United States and Russia should negotiate a compromise in order to settle the Berlin crisis.

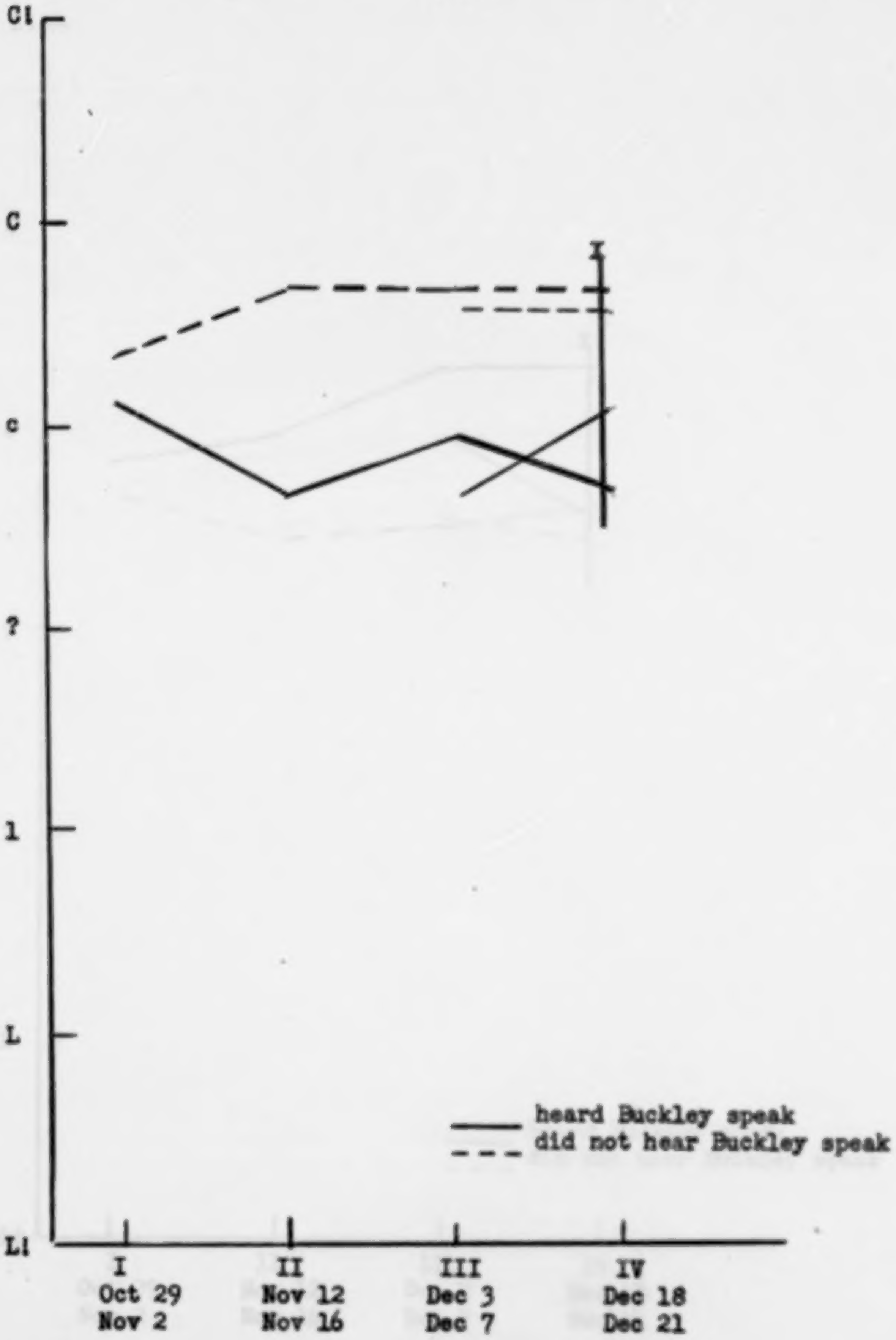
C (B) The United States should tear down the wall separating East and West Berlin.



ITEM 14

L (A) The United States should cease supporting the government of Chiang Kai-shek in Formosa.

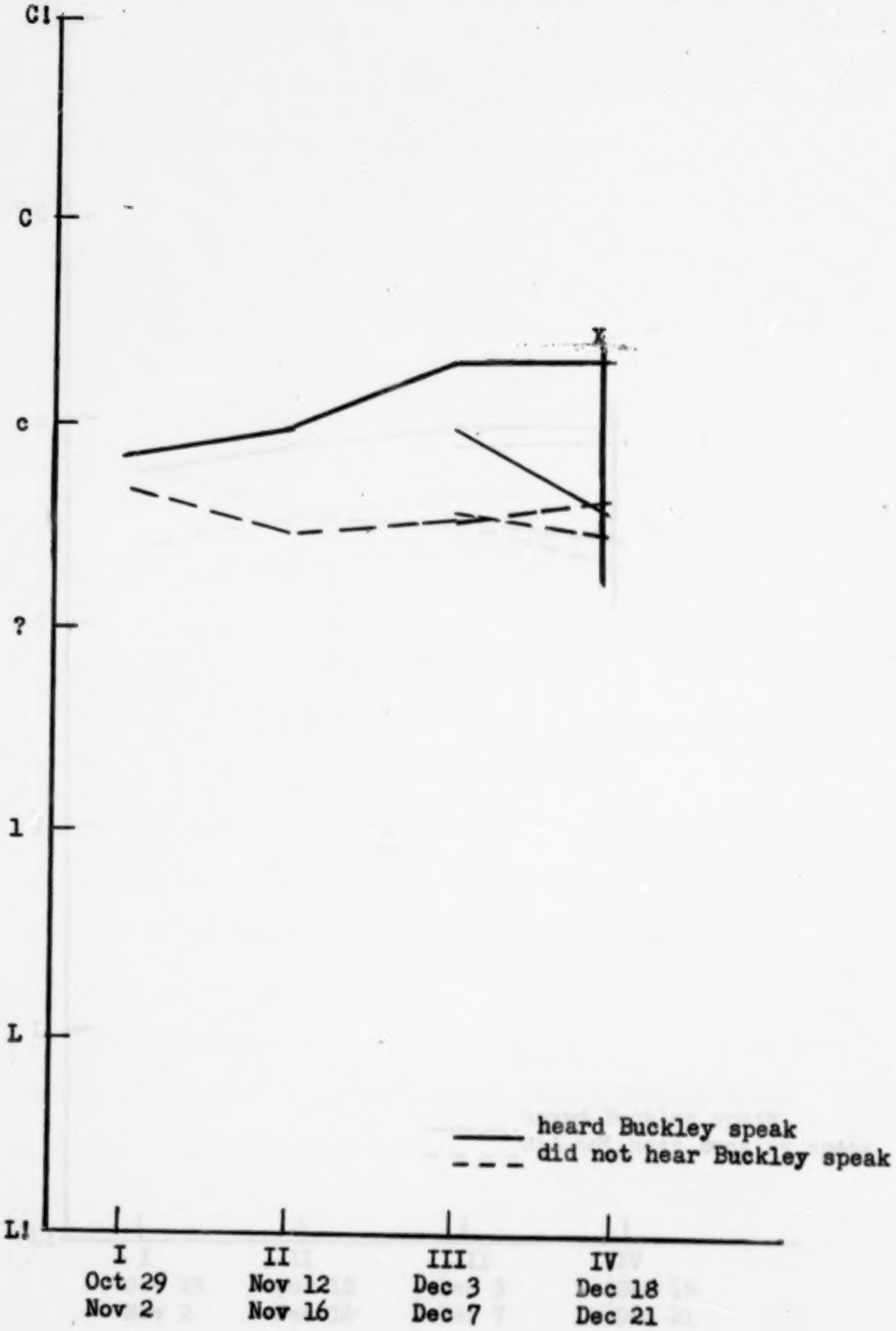
C (B) The United States should continue to support the government of Chiang Kai-shek.



ITEM 15

L (A) Membership in a labor union as a condition of employment is a desirable requirement for skilled and unskilled workers.

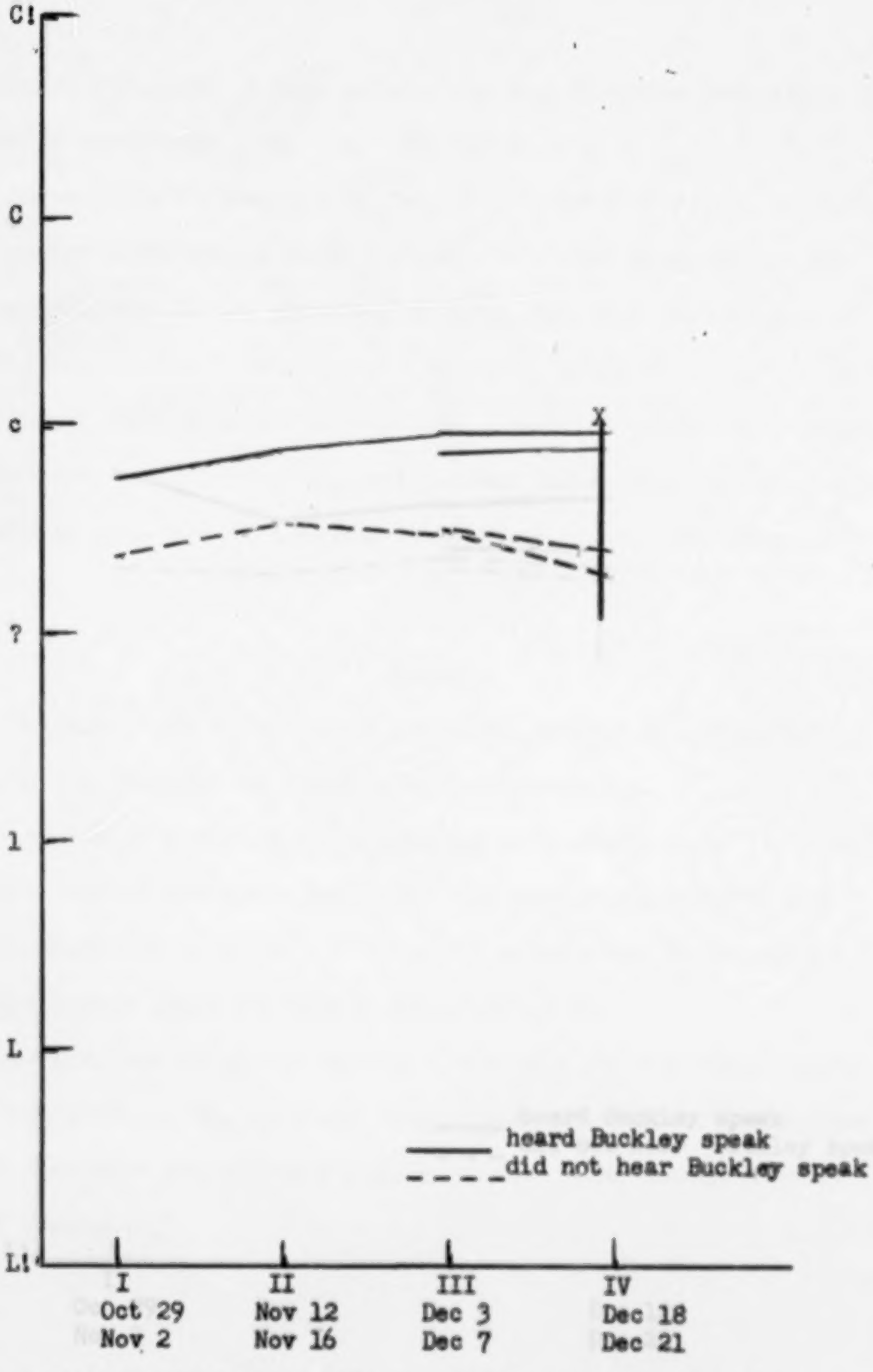
C (B) No man should be compelled to join a union to secure or keep a job.



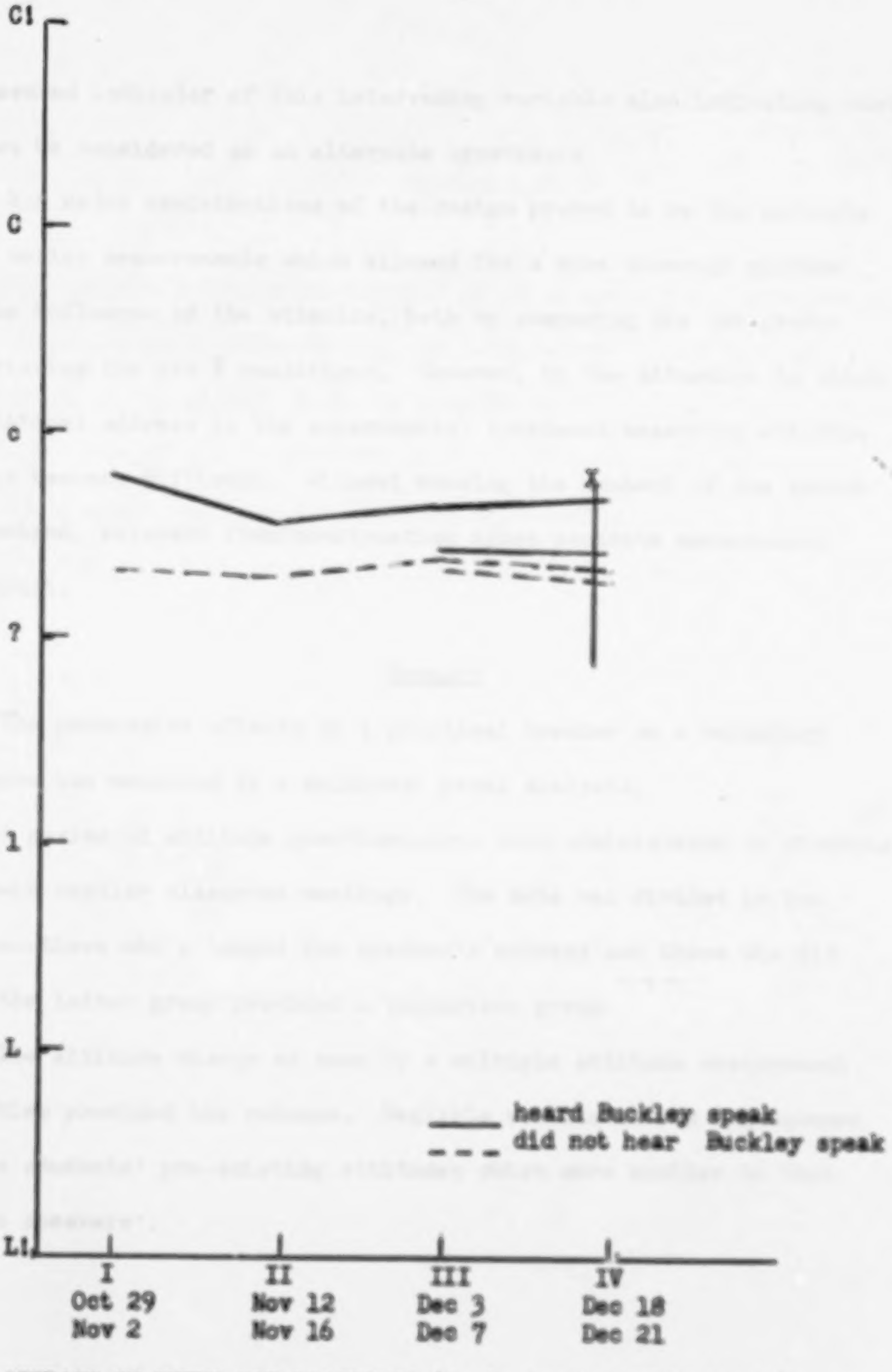
ITEM 16

C (A) Students who receive scholarships granted under the National Defense Education Act should be required to sign a loyalty oath.

L (B) Students who receive scholarships granted under the National Defense Education Act should NOT be required to sign a loyalty oath.



AVERAGE OF DISCUSSED ITEMS (1, 2, 4, 7, 10, 11, 13, 14)



AVERAGE OF ITEMS NOT DISCUSSED (3, 5, 6, 8, 9, 12, 15, 16) a measured indicator of this intervening variable also indicating whether it can be considered as an alternate hypothesis.

The major contributions of the design proved to be the multiple time series measurements which allowed for a more thorough picture of the influence of the stimulus, both by comparing the two groups and viewing the pre X conditions. However, in the situation in which a political address is the experimental treatment measuring attitude change becomes difficult. Without knowing the content of the speech beforehand, relevant item construction makes accurate measurement difficult.

Summary

The persuasive effects of a political speaker on a voluntary audience was measured by a multiwave panel analysis.

A series of attitude questionnaires were administered to students in their regular classroom meetings. The data was divided in two groups--those who attended the speaker's address and those who did not; the latter group provided a comparison group.

The attitude change as seen by a multiple attitude measurement over time provided the outcome. Neglible results were a consequence of the students' pre-existing attitudes which were similar to that of the speakers'.

References

Campbell, Donald T. From description to experimentation: interpreting trends as quasi-experiments. In C. W. Harris (Ed.) Problems in measuring change. Madison, Wis: Univer. Wisconsin Press, 1963.

Campbell, Donald T. & Stanley Julian C. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.) Handbook of research on teaching. Rand McNally & Co., 1963.

Appendix A

STUDENT NUMBER:	____________________	Freshman	____	Sophomore	____
		Junior	____	Senior	____
MAJOR FIELD:	____________________	Age	____		

Trends in Current Opinions

Each item in this survey consists of two alternatives, A and B, between which you are asked to choose. Please indicate your choice by CIRCLING the letter that best exemplifies your attitude or opinion. The explanation of the series of letters is as follows:

A!.......... statement A is strongly preferred over statement B.
A........... statement A is preferred over statement B.
a........... statement A is somewhat preferred over statement B.
?........... I cannot decide between A and B.
b........... statement B is somewhat preferred over statement A.
B........... statement B is preferred over statement A.
B!.......... statement B is strongly preferred over statement A.

Please make sure that you indicate your attitude toward each item.

You are asked to list your student number in the space provided at the top of this sheet. Please do NOT sign your name.

Your willingness to cooperate in this project is deeply appreciated.
THANK YOU!

This survey is being sponsored by the
Department of Psychology, Northwestern University,
Evanston, Illinois

1. (A) The United States should cease giving aid to Communist Yugoslavia and Poland.
   (B) United States aid to Communist Yugoslavia and Poland is essential at the present time.

   A! A a ? b B B!

2. (A) The Supreme Court violated constitutional provisions for states' rights by compelling Southern schools to integrate their classes.
   (B) The Supreme Court decision regarding the integration of public schools was legally and morally justified.

   A! A a ? b B B!

3. (A) The United States would be unwise to withdraw diplomatic recognition from Communist countries.
   (B) The United States should withdraw diplomatic recognition from all Communist countries, including Russia.

   A! A a ? b B B!

4. (A) The United States should depose Fidel Castro of Cuba.
   (B) It would be unwise for the United States to depose Castro.

   A! A a ? b B B!

5. (A) The Federal government should extend financial aid to public schools.
   (B) Each locality should support its own public schools without any Federal assistance.

   A! A a ? b B B!

6. (A) The Federal government is completely justified in placing a 90% tax on incomes over $100,000.
   (B) The Federal government has a right to claim an equal percentage of each man's wealth, and no more.

   A! A a ? b B B!

7. (A) Katanga should not be compelled to submit to the Congolese government.
   (B) Katanga is in no way justified in separating itself from the legitimate government of the Congo.

   A! A a ? b B B!

8. (A) The granting of independence to various African nations is desirable.
   (B) A large number of African nations are completely incapable of governing themselves.

   A! A a ? b B B!

9. (A) Any government program to provide free medical care for the aged should be opposed.
(B) A government program of free medical care for the aged is obviously necessary.

A! A a ? b B B!

---

10. (A) Red China should be admitted to the United Nations.
(B) Red China should NOT be admitted to the United Nations.

A! A a ? b B B!

---

11. (A) The Supreme Court decision regarding prayers in public schools distorts the meaning of the United States Constitution.
(B) The Supreme Court decision prohibiting the saying of prayers in public schools was morally and constitutionally justified.

A! A a ? b B B!

---

12. (A) The United States should not commit itself to any trade agreements with the Soviet Union.
(B) The United States should liberalize its restrictive trade policy vis-a-vis the Soviet Union.

A! A a ? b B B!

---

13. (A) The United States and Russia should negotiate a compromise in order to settle the Berlin crisis.
(B) The United States should tear down the wall separating East and West Berlin.

A! A a ? b B B!

---

14. (A) The United States should cease supporting the government of Chiang Kai-shek in Formosa.
(B) The United States should continue to support the government of Chiang Kai-shek.

A! A a ? b B B!

---

15. (A) Membership in a labor union as a condition of employment is a desirable requirement for skilled and unskilled workers.
(B) No man should be compelled to join a union to secure or keep a job.

A! A a ? b B B!

---

16. (A) Students who receive scholarships granted under the National Defense Education Act should be required to sign a loyalty oath.
(B) Students who receive scholarships granted under the National Defense Education Act should NOT be required to sign a loyalty oath.

A! A a ? b B B!

Appendix B

Tabulation for Series of Two Waves

x...........heard Buckley speak
z...........did not hear Buckley speak
R...........reject
D...........divisor; in determining the means, the divisor is 18 for the x group, and 88 for the z group, unless otherwise noted

	A!						B!				
	1	2	3	4	5	6	7	R	D	Sum	Means
Item 1 (x)											
Wave 3	2	6	4	1	2	2	1			59	3.28
4	2	5	3	0	5	2	1			65	3.61
Item 1 (z)											
Wave 3	23	10	14	7	14	13	6	1	87	303	3.48
4	14	11	11	6	17	20	9			361	4.10
Item 2 (x)											
Wave 3	2	1	1	0	3	3	8			96	5.33
4	2	3	1	0	1	3	8			90	5.00
Item 2 (z)											
Wave 3	9	12	5	1	1	24	36			453	5.15
4	7	9	5	2	11	24	30			457	5.19
Item 3 (x)											
Wave 3	7	6	2	0	1	2	0			42	5.67
4	7	6	4	0	1	0	0			36	6.00
Item 3 (z)											
Wave 3	20	33	21	5	3	4	2			222	5.48
4	22	33	22	1	7	3	0			211	5.60
Item 4 (x)											
Wave 3	4	1	4	0	5	4	0			67	3.72
4	4	2	3	0	5	4	0			66	3.67
Item 4 (z)											
Wave 3	22	8	18	7	12	13	8			314	3.57
4.	20	8	14	4	16	17	9			339	3.85
Item 5 (x)											
Wave 3	3	3	8	1	0	1	1	1	17	50	5.06
4	2	3	5	2	2	3	1			66	4.33
Item 5 (z)											
Wave 3	24	21	19	3	6	8	7			262	5.02
4	16	22	17	6	8	11	8			297	4.62
Item 6 (x)											
Wave 3	1	3	3	0	4	3	4			82	3.44
4	1	3	3	3	3	1	4			77	3.72
Item 6 (z)											
Wave 3	8	15	17	5	15	14	14			366	3.84
4	8	11	20	7	12	16	14			372	3.77

	1	2	3	4	5	6	7	R	D	Sum	Means
Item 7 (x)											
Wave 3	4	3	3	3	2	2	1			60	3.33
4	8	2	2	3	2	0	1			47	2.61
Item 7 (z)											
Wave 3	7	11	17	22	11	12	7	1	87	344	3.95
4	8	9	21	18	10	15	5	2	86	336	3.91
Item 8 (x)											
Wave 3	2	1	1	1	3	6	4			90	3.00
4	1	1	2	0	5	4	5			93	2.83
Item 8 (z)											
Wave 3	4	6	15	1	17	24	21			441	2.99
4	3	9	13	2	15	24	22			441	2.99
Item 9 (x)											
Wave 3	2	3	4	1	1	3	3	1	17	68	4.00
4	2	3	2	2	6	1	2			72	4.00
Item 9 (z)											
Wave 3	7	2	10	5	23	26	14	1	87	430	4.94
4	5	8	4	4	27	30	10			434	4.93
Item 10 (x)											
Wave 3	0	0	0	1	0	2	15			121	1.28
4	1	0	0	0	1	3	13			115	1.61
Item 10 (z)											
Wave 3	4	3	1	2	5	13	58	2	86	530	1.84
4	6	1	3	1	6	11	60			537	1.90
Item 11 (x)											
Wave 3	8	4	1	1	1	3	0			46	2.56
4	7	3	4	2	2	0	0			43	2.39
Item 11 (z)											
Wave 3	44	14	15	6	5	2	2			192	2.18
4	32	25	12	4	11	3	1			214	2.43
Item 12 (x)											
Wave 3	5	4	4	3	2	0	0			47	2.61
4	5	4	1	5	2	1	0			52	2.89
Item 12 (z)											
Wave 3	16	25	15	6	15	7	3	1	87	273	3.14
4	16	20	12	11	17	8	3	1	87	290	3.33

	1	2	3	4	5	6	7	R	D	Sum	Means
Item 13 (x)											
Wave 3	1	2	4	2	2	0	6	1	17	77	3.47
4	0	1	5	2	3	4	3			85	3.28
Item 13 (z)											
Wave 3	13	17	24	5	8	8	12	1	87	311	4.43
4	14	20	23	7	10	7	6	1	87	285	4.72
Item 14 (x)											
Wave 3	1	0	0	0	1	6	10			112	1.78
4	1	0	1	0	1	8	7			106	2.11
Item 14 (z)											
Wave 3	4	7	4	7	6	23	35	2	86	471	2.52
4	1	2	8	7	10	26	32	2	86	487	2.34
Item 15 (x)											
Wave 3	2	1	5	1	0	3	6			83	3.39
4	3	1	0	0	4	2	7	1	17	86	2.94
Item 15 (z)											
Wave 3	1	10	6	1	11	20	38	1	87	484	2.44
4	2	10	7	0	10	18	40	1	87	481	2.47
Item 16 (x)											
Wave 3	6	3	4	0	1	3	1			54	3.00
4	6	2	1	2	0	5	1	1	17	58	3.41
Item 16 (z)											
Wave 3	27	8	15	9	11	5	13			300	3.41
4	21	13	13	4	17	14	6			313	3.56

Tabulation for Series of Four Waves

x..........heard Buckley speak
z..........did not hear Buckley speak
R..........reject
D..........divisor; in determining the means, the divisor is 17 for the x group, and 71 for the z group, unless otherwise noted

	A!						B!				
	1	2	3	4	5	6	7	R	D	Sum	Means
Item 1 (x)											
Wave 1	7	2	2	1	3	1	1			49	2.88
2	7	1	0	1	5	3	0			56	3.29
3	8	0	4	1	1	2	1			48	2.82
4	6	1	2	1	4	3	0			56	3.29
Item 1 (z)											
Wave 1	18	8	13	6	12	8	6			247	3.48
2	11	17	14	4	11	7	6	1	70	242	3.46
3	15	13	11	3	17	6	6			249	3.51
4	14	15	8	4	14	8	7	1	70	251	3.59
Item 2 (x)											
Wave 1	2	1	0	1	2	3	8			92	5.41
2	1	1	0	1	3	3	8			96	5.65
3	2	2	1	1	3	1	7			83	4.89
4	4	2	0	0	4	2	5			75	4.41
Item 2 (z)											
Wave 1	3	0	3	0	6	13	46			442	6.23
2	4	1	6	0	1	18	41			424	5.97
3	4	3	5	2	5	15	37			407	5.73
4	1	6	6	3	5	9	41			409	5.76
Item 3 (x)											
Wave 1	2	5	5	1	0	1	3			58	4.59
2	2	5	4	1	1	2	2			59	4.53
3	1	4	4	1	2	5	0			65	4.18
4	2	3	4	1	2	3	2			66	4.24
Item 3 (z)											
Wave 1	25	27	10	3	1	1	3	1	70	153	5.81
2	19	23	22	1	4	1	1			168	5.63
3	20	19	19	1	4	3	4	1	70	185	5.36
4	17	23	16	1	7	5	2			194	5.27
Item 4 (x)											
Wave 1	3	5	3	1	1	4	0			55	3.24
2	6	2	4	1	3	0	0	1	16	41	2.56
3	5	3	4	2	2	1	0			47	2.76
4	5	4	2	2	3	1	0			48	2.82
Item 4 (z)											
Wave 1	27	8	10	3	4	9	10			229	3.23
2	28	13	10	0	5	8	7			206	2.90
3	22	8	8	5	13	7	8			245	3.45
4	18	8	10	5	11	13	6			259	3.65

	1	2	3	4	5	6	7	R	D	Sum	Means
Item 5 (x)											
Wave 1	1	6	3	1	2	1	3			63	4.29
2	2	4	1	2	1	4	3			71	3.82
3	2	5	3	2	0	0	5			64	4.24
4	2	3	3	2	1	0	6			72	3.76
Item 5 (z)											
Wave 1	12	23	10	3	6	11	6			238	4.65
2	13	18	13	5	9	6	7			238	4.65
3	6	18	20	4	8	8	6	1	70	248	4.46
4	7	20	16	4	6	12	5	1	70	248	4.46
Item 6 (x)											
Wave 1	1	2	3	1	4	3	3			77	3.47
2	3	2	4	2	1	3	2			64	4.24
3	1	4	3	2	3	3	1			66	4.12
4	2	2	3	2	3	2	3			71	3.82
Item 6 (z)											
Wave 1	11	6	17	9	7	10	9	2	69	268	4.12
2	13	7	12	8	4	15	12			289	3.93
3	12	11	10	7	4	11	15	1	70	283	3.96
4	9	12	11	11	4	8	15	1	70	283	3.96
Item 7 (x)											
Wave 1	3	3	4	3	0	2	2	1	16	57	3.56
2	3	3	4	3	1	1	2			58	3.41
3	3	2	5	3	1	0	3			60	3.53
4	5	1	3	3	2	1	2			58	3.41
Item 7 (z)											
Wave 1	7	10	7	25	7	8	7			280	3.94
2	5	12	7	23	10	8	6			282	3.97
3	7	8	10	18	14	7	6	1	70	279	3.99
4	7	12	8	22	7	9	6			274	3.86
Item 8 (x)											
Wave 1	0	0	3	2	5	5	2			86	2.94
2	0	3	1	0	4	5	4			87	2.88
3	0	2	2	0	3	6	4			89	2.76
4	1	1	3	1	5	2	4			81	3.24
Item 8 (z)											
Wave 1	2	7	5	6	11	20	19	1	70	363	2.81
2	1	7	7	4	16	19	17			365	2.86
3	2	8	3	4	16	24	13	1	70	358	2.96
4	3	8	7	4	14	21	13	1	70	343	3.18

	1	2	3	4	5	6	7	R	D	Sum	Means
Item 9 (x)											
Wave 1	3	2	4	0	2	2	4			69	4.06
2	1	2	2	0	5	2	5			83	4.88
3	2	2	3	0	4	1	5			76	4.47
4	4	3	0	1	4	1	4			68	4.00
Item 9 (z)											
Wave 1	7	4	6	2	15	17	20			358	5.04
2	4	8	2	4	16	17	20			364	5.13
3	4	7	2	9	17	17	13	2	69	338	4.76
4	5	6	5	5	22	13	15			345	4.86
Item 10 (x)											
Wave 1	1	0	1	0	0	1	14			108	1.65
2	1	0	0	0	1	2	13			109	1.59
3	1	0	0	0	1	1	14			110	1.53
4	0	0	0	1	0	3	13			113	1.35
Item 10 (z)											
Wave 1	10	3	1	0	2	7	48			407	2.27
2	7	1	2	2	2	9	48			423	2.04
3	6	3	2	1	1	9	48	1	70	417	2.04
4	8	3	2	2	2	7	46	1	70	402	2.34
Item 11 (x)											
Wave 1	8	3	3	1	1	0	1			39	2.29
2	6	4	3	1	1	1	1			45	2.65
3	5	3	4	1	1	1	2			52	3.06
4	3	3	5	1	2	1	2			58	3.41
Item 11 (z)											
Wave 1	29	14	12	3	7	2	3	1	70	173	2.44
2	28	12	11	4	8	6	2			191	2.69
3	31	14	6	2	7	6	4	1	70	184	2.63
4	29	18	3	4	8	7	1	1	70	179	2.52
Item 12 (x)											
Wave 1	10	2	3	1	0	0	0	1	16	27	1.69
2	10	3	2	0	1	0	1			34	2.00
3	8	5	2	1	0	1	0			34	2.00
4	11	1	3	1	0	1	0			32	1.88
Item 12 (z)											
Wave 1	18	22	14	4	7	3	2	1	70	187	2.67
2	19	17	9	6	11	5	4			217	3.06
3	15	13	17	6	10	5	3	2	69	217	3.14
4	17	11	15	8	9	6	5			232	3.27

	1	2	3	4	5	6	7	R	D	Sum	Means
Item 13 (x)											
Wave 1	2	3	1	1	3	3	4			76	3.53
2	0	3	4	0	2	4	4			80	3.29
3	0	2	4	0	0	7	4			86	2.94
4	0	2	4	1	1	4	5			84	3.06
Item 13 (z)											
Wave 1	10	10	15	4	12	12	7	1	70	272	4.11
2	6	14	14	4	11	13	9			288	3.94
3	6	14	9	6	16	3	15	2	69	288	3.83
4	6	10	9	7	16	7	16			315	3.56
Item 14 (x)											
Wave 1	1	1	2	0	2	5	6			91	2.65
2	1	0	2	0	1	7	6			96	2.35
3	1	1	2	1	3	2	7			89	2.77
4	2	0	1	0	6	2	6			89	2.77
Item 14 (z)											
Wave 1	2	3	2	5	13	12	32	2	69	395	2.28
2	1	3	5	6	9	14	33			406	2.28
3	1	4	1	7	9	17	31	1	70	404	2.23
4	2	3	6	7	9	15	29			392	2.48
Item 15 (x)											
Wave 1	1	3	0	1	3	2	7			87	2.88
2	3	2	0	0	5	2	5			79	3.35
3	2	2	2	0	1	3	7			84	3.06
4	1	4	2	1	0	3	6			79	3.35
Item 15 (z)											
Wave 1	1	11	5	3	8	12	31			379	2.66
2	3	3	4	3	9	20	29			401	2.35
3	3	5	2	3	11	13	33	1	70	395	2.36
4	1	6	4	4	8	17	31			400	2.37
Item 16 (x)											
Wave 1	5	3	1	4	1	2	1			54	3.18
2	8	2	0	1	3	1	2			51	3.00
3	6	5	1	1	2	1	1			46	2.71
4	6	5	1	1	2	1	1			46	2.71
Item 16 (z)											
Wave 1	20	15	5	9	6	7	9			236	3.32
2	17	14	7	9	5	10	9			250	3.52
3	21	9	9	7	7	6	11	1	70	242	3.46
4	25	8	6	8	5	9	10			240	3.38

Invited Therapy in an Archival or Institutional-Records Setting: with Comments on the Problem of Turndowns[1]

Donald T. Campbell

Northwestern University

Research in the natural setting upon the effectiveness of remedial procedures such as psychotherapy is notoriously difficult and has been unsatisfactory for two reasons:

First difficulty. Since the usual therapy patients are self-selected, it is nearly impossible to achieve an appropriate no-therapy control group. If non-applicants for therapy are used, matching produces regression effects confounded with treatment effects, and covariance analyses are likewise inappropriate because the assumption of homogeneity of regression becomes patently implausible. If therapy applicants are randomly divided into experimental and control groups, the frustrating postponement of treatment for the control group makes them an inappropriate comparison group. Differential drop-out of the two groups and the awkwardness of matching control group delays to extended and variable-length therapy periods for experimental patients add further troubles. For some therapies, and in particular for supplementary fixed-term remedial courses in the educational setting, delayed-treatment control group designs are satisfactory and valuable, but without discouraging their increased use, exploration of alternate models seems very desirable.

---

[1] The preparation of this paper was supported in part by Project C-998, Contract 3-20-001, with the Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act. The development of the ideas has profited from extensive discussions with Professors Robert I. Watson and Herbert C. Quay.

Research in psychotherapy frequently avoids the delayed-treatment control group by abandoning the no-treatment comparison entirely, and limiting itself to a comparison among therapies, with random assignment of all applicants to one or another of the treatments being compared. If no difference is found between the therapies, it is apt to be concluded that both were equally effective, rather than that neither had any effect, although usually no conclusive comparisons to rule out the latter interpretation are at hand. Even if one of the comparison treatments is a placebo therapy, the comparison is still difficult, because placebo treatments are often in themselves genuinely therapeutic. While placebo therapy groups should be used, they do not eliminate the need for a no-therapy comparison group.

Second difficulty. _Measures of adjustment which are non-reactive and independent of potential self-deception by patient and therapist are usually lacking_. Most therapy research has to depend upon measures of effect based upon the judgments of the therapist or the patient, of such a nature that the judge is fully aware of the implications of his report. Such measures would include direct ratings, personality or adjustment inventories, Q sorts, sentence-completion tests and the like. Even when measures such as the Rorschach are used, for which the evaluative implications are not obvious to the patient, the measurement process is a conspicuous, artifical procedure intruded for an obvious purpose. Further such measures are apt to be of dubious relevance to the performance of everyday life which led to the patient's referral. When more objective criteria are introduced, such as ratings by acquaintances not involved in the therapeutic process (perhaps even unaware that the person is under therapy) the measurement process is nonetheless a highly conspicuous and artificial intrusion into the normal course of events, and apt to be reactive insofar as the raters become curious and try to figure out the reasons for it, or communicate about it to the patient as one friend to another, etc. How much better it would be if some of the measures at least were a natural part of the situation in which the patients and the controls normally lived, and were thus non-reactive and independent of the experimental program.

Invited therapy in a setting providing regular measurements. The purpose of this paper is to advocate the joint use of two procedures which have already been partially applied in previous research, and to elaborate on a serious problem of analysis which results. The first feature is for the researcher to invite persons thought to need it into therapy. By so doing, a randomized subset of those in need can be designated as a control group without their awareness.

Second, it is proposed that situations for such experimental invitation be sought out which become "natural laboratories" through providing regular ongoing measurements which are a natural and integral part of the situation. Since the social scientists' usual use of such records has been retrospective, these have been called "archival" measures. Some settings provide these of such a nature that they are relevant to the performance adequacy of individuals. For research on individual therapies, these become the natural laboratories. To make the discussion more concrete, let psychotherapy be the representative of remedial treatments in the discussion that follows.

Examples of such archival settings would include school systems with frequent marking periods, absences and participation records; factories keeping records of units produced per person; sales records as for life insurance, etc. Such settings would be selected on the basis of the quality of records available and relevance to likely effects from therapy. Within such a setting, supervisory persons would be asked to designate a set of persons whose maladjustments are such that they impair their performance on the records available. Thus persons whose neuroses lead to too high grades, too great sales industry, insensitivity to social pressures against rate-busting, etc. would be excluded or treated as a separate problem. The resulting pool of appropriate candidates would be divided by randomization (possibly after matching) into a control group not told of its status, and an experimental group invited to participate in psychotherapy experimentally provided.

Problems in comparability still remain if not all those invited accept. However, a statistically acceptable, though conservative, solution is always open, i.e., to compare the total invited group (accepting plus non-accepting) with the total non-invited group. Even though the non-accepters dilute the effect, the total of the invited group should show greater mean gain in effectiveness. Possible tests of greater precision will be discussed below.

Ideally the financing of the research would be such that those accepting psychotherapy would receive it, at the project's cost, from professional psychotherapists.

The major goal of the study would be to compare the standard therapies available in a large metropolitan community with the absence of any therapy. For only one year the number treated would have to be considered as a single treatment group, but over a five year period enough cases would be accumulated to make possible comparisons among the schools or styles of therapy available locally in private practice. Therapists of diverse points of view, psychoanalytic, eclectic, Adlerian, Rogerian and others are typically available.

Note the difference in approach between this mode of research on psychotherapy in practice, and the usual situation. In the usual research, the scientist must ask the therapist's cooperation, and intrudes into the process with measurement procedures foreign to the usual therapeutic situation. This intrusion provides ample moral ground for the therapist's refusal. And even if cooperation is achieved from some therapists it is from a highly select subgroup. Further, any comparisons between types of therapy must presume an improbable equivalence of patient populations received by different therapists. On the other hand, in the setting of this proposal the therapist is approached only on a client-by-client basis and any therapist offering treatment within the appropriate age-sex-problem area should be accessible.

Modes of Data Analysis

Conventional methods of statistical analysis could be used but certain modes of data analysis particularly appropriate to the design in question should be specified. A conservative analysis of data using total-invited and control groups should be made, with t tests, analysis of variance or analysis of covariance of post-test on pretest scores, as appropriate to the aspect in question. If, of those invited, almost all accept therapy, these tests will be as appropriate or powerful as any. By paying for the therapy and by selecting invitees in enough trouble to recognize this themselves, the nonacceptance rate in the experimental group might be as low as 10% or 20%. However, this cannot be counted on, and non-acceptance rates as high as 70% or 80% might be encountered. A possible mode of focusing the analysis upon those who get therapy, without introducing self-selection and regression biases, is discussed in the section on "statistics for biased partial exposure designs" which follows immediately. The subsequent section on "time-series analysis" provides a quasi-experimental perspective applicable to both groups and individual cases. The three experimental and quasi-experimental designs discussed have been selected as those most appropriate for psychotherapy research after a review of some 16 such designs (Campbell, 1957; Campbell & Stanley, 1963).

1. Statistics for biased partial exposure designs. Let us suppose 10% of a high school class judged as being able to profit from psychotherapy is divided randomly into an invited group and a non-invited control group. Of the invited group, only 30% accept the invitation and get exposed to the therapy. For all students, gain scores are computed for grades one year later. Let us suppose that these gain scores are distributed as shown in Fig. 1. In this graph, the *N*s represent students not exposed to the remedial treatment. The *X*s represent those few who took therapy. If the outcome were as shown, visual inspection would leave us certain that the treatment was effective. Our problem is to find tests of significance appropriate to this high-mortality condition. We can first note that to show that the 15 treated students gain more than the uninvited control group and the non-volunteering invited is essential (in that if this were *not* so we would certainly not retain the belief in the efficacy of therapy) but not sufficient.

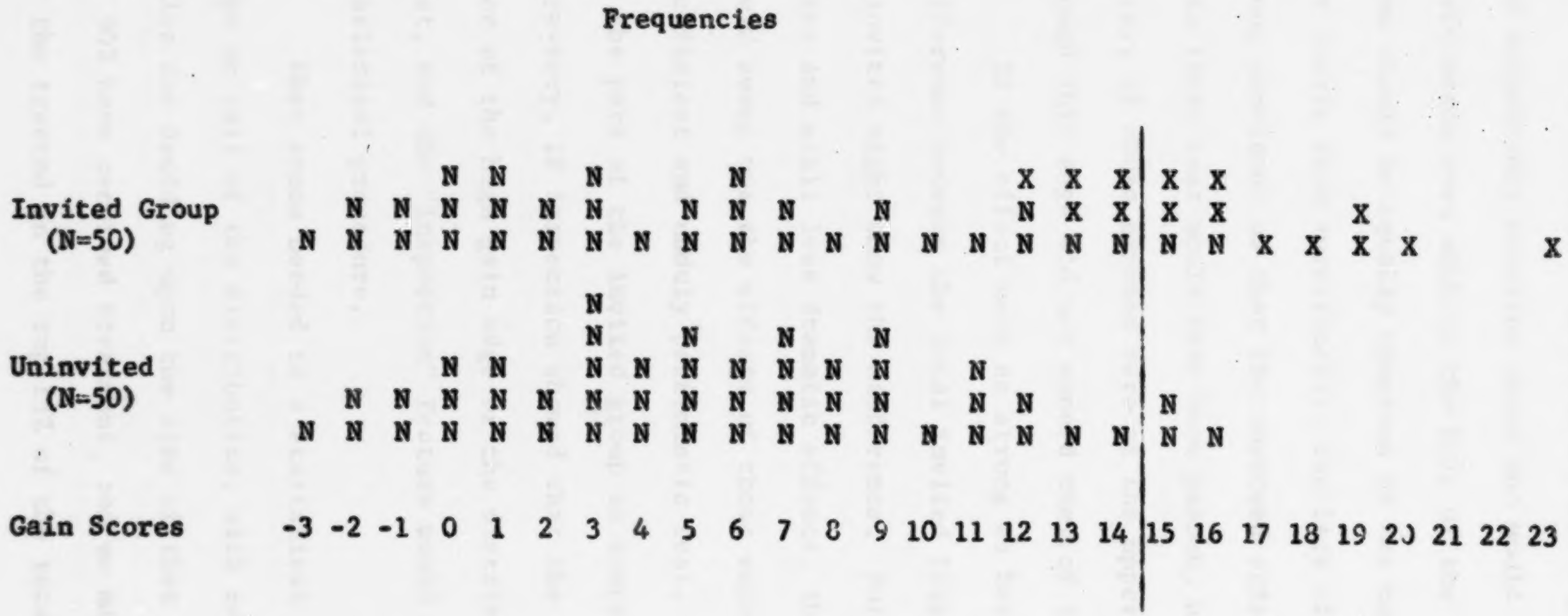


Fig. 1. Hypothetical results from an Invited Remedial Treatment experiment, in which 30% of those initiated received the treatment. (X = treated, N = nontreated)

It is not sufficient inasmuch as this exposed subgroup is highly self-selected and undoubtedly contains those who would have done the most to improve their performance even without the help of the remedial treatments. But such persons should be equally numerous in the control group, and even though we cannot locate them specifically, the lack of equally high gainers in the control group convinces us that the observed effect is more than just this selection. This first test would have been passed, however, without this being the case (i.e., if those exposed were at the upper edge of the invited group, even though this edge did not exceed that of the uninvited).

If the effect were as strong as here diagrammed, a t test of the mean difference between the total invited (exposed plus nonexposed) and the total uninvited might show the difference. But in cases of still smaller volunteer rates and still less dramatic effects, the dead weight of the nonvolunteers could swamp out the effects of those exposed, making such a total-group t an inefficient and unduly pessimistic test. A significantly greater variance on the part of the invited group as compared with the uninvited would be confirmatory, if inspection showed that the greater variance was due to extension of the high gain edge of the distribution. This would still be a dull test, and the "inspection" feature would not be formally represented in the statistical procedure.

What seems needed is a statistical procedure focusing upon the single edge or tail of the distribution, with some a priori (non-opportunistic) rules for deciding upon the size of that tail. In the present instance, 30% of 50% have received treatment, and we might therefore look to the presence of the treated in the top 15% of the total invited-plus-uninvited. A cut has been inserted at this point, and we find that of these 15, 12 come from the invited group (2 non-exposed, 10 exposed). An appropriate test of significance would be a chi-square test based upon the four-celled contingency table of invited-noninvited (50-50) by top 15%--bottom 85%, with one degree of freedom. In the illustrative instance, the computations become:

	Bottom 85%	Top 15%	Total
Invited	38 (42.5)	12 (7.5)	50
Uninvited	47 (42.5)	3 (7.5)	50
Total	85	15	100

The Chi-square total is 6.4, $.01 < p < .02$.

The degree to which this achieves a gain through focusing upon one extreme edge of the distribution can be shown by comparing the Chi-square of 6.4 obtained above with that produced by a median split of the same gain scores:

	Bottom 50%	Top 50%	Total
Invited	22 (25.0)	28 (25.0)	50
Uninvited	28 (25.0)	22 (25.0)	50
Total	50	50	100

Chi-square = 1.44, $p > .10$

It is possible that more parametric tests focused on a single edge could be developed through examining the distribution of the top 15% of extreme scores in terms of their individual p values based upon the mean and standard deviation of the pooled distribution.

2. Time-series data in quasi-experimental analysis. The setting chosen makes possible a type of analysis, common to medical research and close to the common-sense interpretation of the effects of specific events upon frequently seen acquaintances. This form of analysis, however, is rare in psychotherapy research or in psychological research more generally.

Let us suppose that one of our therapy cases shows the trend of Fig. 2 in average grade during several years in high school. If such data were obtained in an insulated biological laboratory, in which the experimenter felt sure no other variable had intruded contemporaneously with the introduction of the experimental treatment, the experimenter would have little hesitancy in concluding that the treatment had had an effect. While the biologist's experimental isolation, control and accompaning assurance about the absence of extraneous variables would be lacking, it would nonetheless seem reasonable to consider a comparable interpretation. The possible rival hypotheses can be examined for their plausibility one by one. Thus, a change of teachers, or a change of course content, or a change in the patient's residence or family constellation all could be examined and possibly ruled out. For these purposes strength is added through the presence of control cases in the same classrooms, and the presence of several separate course records, making possible an evaluation of the reliability of the trend.

What we wish to call attention to is the advantage of the time-series of measurements over a single before and a single after measure. The interpretation of the immediate rise at the onset of therapy could be of the same magnitude as shown, and be completely unconvincing as evidence of effect under several conditions, e.g., if the pre-therapy trend had been steadily

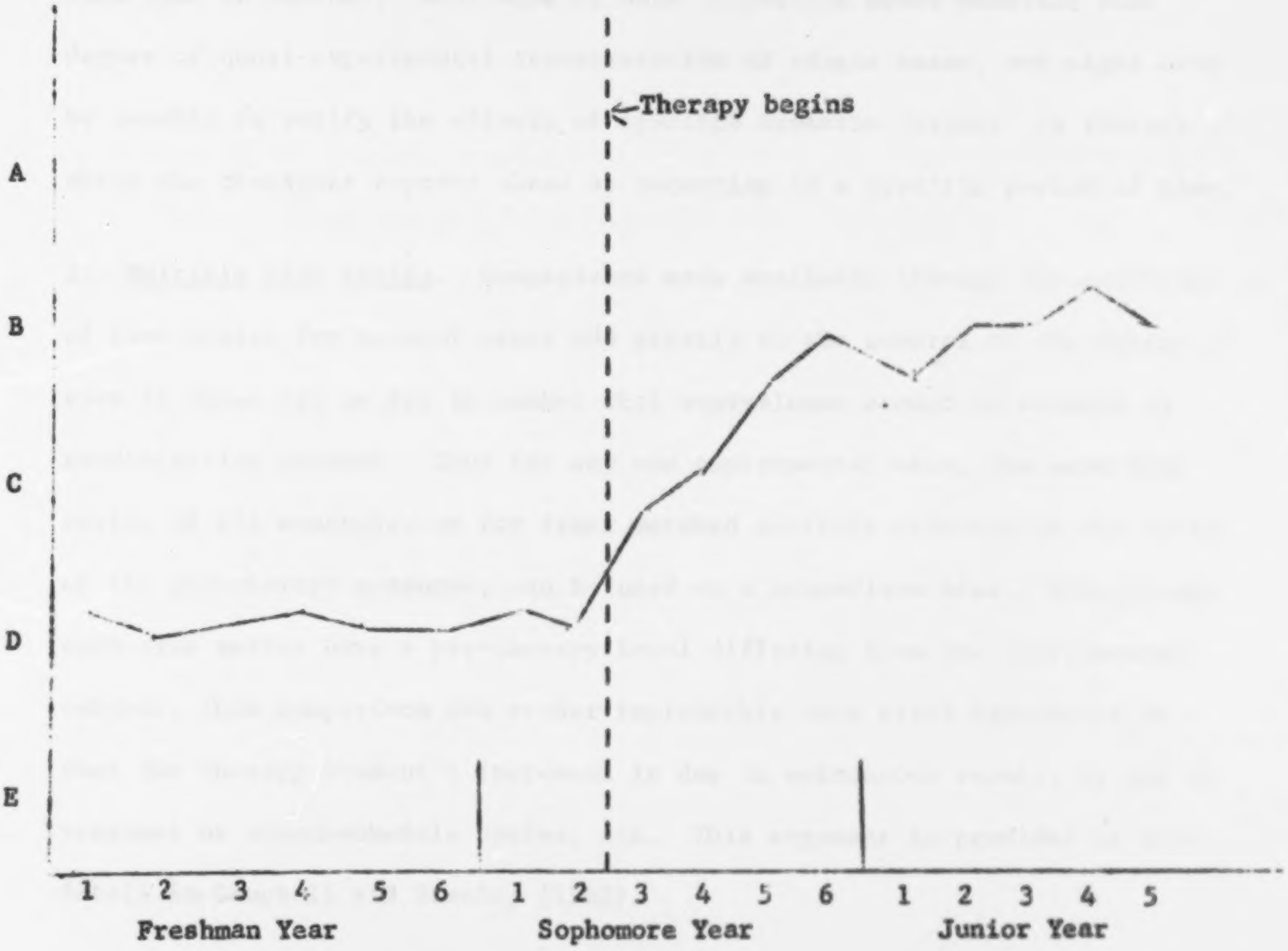


Fig. 2. Hypothetical Trend in one student's grades over 6 marking periods per school year for three years.

upward, or if it had been so eratic that there had been several equally abrupt changes prior and subsequently. Our interpretation of the rise is thus helped a great deal by the evidence of the nature of the trends for some time in advance. This mode of data inspection makes possible some degree of quasi-experimental interpretation of single cases, and might even be usuable to verify the effects of specific dramatic "events" in therapy where the therapist reports these as occurring in a specific period of time.

3. Multiple time series. Comparisons made available through the existence of time series for control cases add greatly to the control of the design even if these are so few in number that equivalence cannot be assured on randomization grounds. Thus for any one experimental case, the mean time series of all controls, or for fewer matched controls selected on the basis of the pre-therapy measures, can be used as a comparison base. Even though such time series have a pre-therapy level differing from the experimental subject, this comparison can render implausible such rival hypotheses as that the therapy student's increment is due to extraneous events, or due to seasonal or school-schedule cycles, etc. This argument is provided in more detail in Campbell and Stanley (1963).

Summary

It is recommended that research on therapy comparing it with no therapy be undertaken in settings in which institutional records provide repeated individual performance measures. In such settings, an appropriate needy experimental group would be invited to accept psychotherapy while a randomized equivalent control was not. A major problem in such a design is a large percentage of the invited who fail to accept, and for this several statistical procedures are suggested. Major advantages are the avoidance of self-selection bias in comparing therapy groups with controls, and the use of relevant life performance measures of a non-reactive nature.

References

Campbell, D. T. Factors relevant to validity of experiments in social settings. Psychological Bulletin, 1957, 54:297-312.

Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. Pages 212-242 in Harris, C. W. (ed.), Problems in measuring change. Madison, Wis.: Univ. Wisconsin Press, 1963.

Campbell, D. T., and Stanley, J. S. Experimental and quasi-experimental designs for research on teaching. Pages 171-246 in Gage, N. L. (ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

EVALUATING AN EXPERIMENTAL COLLEGE PROGRAM WITH INSTITUTIONAL RECORDS: AN INTERIM REPORT*

Sally W. Cassidy, Paule Verdet, Richard H. Schell,
Wayne State University
D. T. Campbell, Northwestern University

In spite of the great research efforts that have been expended on evaluating the effects of educational programs at the college level we have no interpretable evidence that any college has made a difference or that any one college produces an outcome differing from another. The reason why such a devastating statement can be made is that in almost every setting the differential effects of programs are inextricably confounded with the differential selection of those exposed to the programs. Thus it is not enough to show that seniors differ from freshmen in interests, values, and career choices -- any such study is taking place during a period of life characterized by rapid changes for everyone, in college or out, and we have no appropriate "control group" set aside to show which of these changes are specific to the college experience. Nor is it enough to show that alumni from Harvard ten years after graduation are earning twice as much as those from Rocky Mountain State. In terms of family position, point of entry into the economic system, etc., they probably would have had this advantage even without the Harvard training. And even were we to "control" by an examination of those going to Harvard from Rocky Mountain's recruitment region, economic and ability strata, we would not have controlled for that high valuing of intellectual, social and economic mobility which led to the differential choice in the first place. And these are only the first phalanx of difficulties faced in such evaluational efforts. Technical difficulties such as regression artifacts which invalidate all matching, ex post facto and control procedures are too complex for elaboration here. Suffice it to say that it is not from lack of skill or effort that another Ivy League university spent half-a-million dollars under the direction of an eminent statistician trying and failing to evaluate the impact of that school upon its students.

---

*Supported in part by Project C-998, Contract 3-20-001, Educational Media Branch, Office of Education, U. S. Department of Health, Education, and Welfare, under the provisions of Title 7 of the National Defense Education Act; Donald T. Campbell, Northwestern University, Principal Investigator.

This pessimistic description of the current state of our knowledge is provided as a background to explain and to justify, if need be, the tedious and frustrating study here being reported at mid-stage. For we have been grasping at what might be the most interpretable experiment in college education to date. If this best is not very good, if any inferences drawn are jeopardized by numerous plausible rival hypotheses, this is the state of our art. This state can and must be improved. It is our hope that by a full revelation of the imperfections in our design and data, we can provide the understanding and influence the decisions that will make possible better experiments in the future.

Monteith College was established at Wayne State University as an experimental effort to offer an ideal small-college experience to a representative sample of the Wayne student body. Whereas in most such demonstration projects those admitted are a highly selected group, uncomparable with any available previous or contemporary comparison group, for Monteith College, at least for the first few years, an effort to achieve comparability was made by inviting only a "random" portion of Wayne applicants into Monteith. It is this feature, plus the radical contrast between Monteith and the rest of Wayne, which made this seem such an important research opportunity.

The Monteith College experiment is being reported elsewhere in terms of a variety of interview and case study procedures. The present study differs from these in two details: The first is its focus on the process of randomization in establishing a control group of those not invited, and in collecting the same data on these uninvited controls as for the Monteith experimentals.

The second difference is in restricting the evidence of effect to those institutional records, within the university and without, which could reflect alike on the behavior of experimentals and controls. The study was thus also an experiment in the use of such records for research purposes.

Institutional records are potentially one of the most important resources available for the social scientist, and yet he usually fails to use them, preferring to construct new questionnaires for his respondents to fill out. Institutional records have many advantages over questionnaires and interviews.[2] One of these is their "nonreactive" nature. Usual social science research procedures intrude into the respondent's life, creating attitudes, initiating change processes, creating commitment effects, producing self-consciousness, and the like. All of these limit generalization. At the worst, the pretesting may account for all of the observed change, without any real effect from the experimental treatment. At a second level, even if there is an experimental effect, generalization from a pretested to an unpretested population is often unjustified, and it is just this generalization that we want to make. Institutional records, being a part of the "normal" environment and being typical of the situations into which we want to generalize, lack these problems. In addition, we presumably have no refusal rate. Furthermore, they may make it possible for us to go back in time and reconstruct an experimental evaluation of a past social experiment. The present study is a case in point.[3] While institutional records should never supplant other procedures, they should regularly be used to supplement, cross-check, and validate.

---

[2] See E. J. Webb, D. T.Campbell, R. Schwartz and L. Sechrest, *Unobtrusive Measures*. Chicago: Rand McNally, 1966 (in Press).

[3] Another illustration is provided by H. L. Ross & D. T. Campbell, "Time-Series Data in the Quasi-Experimental Analysis of the Connecticut Speeding Crackdown." Mimeographed Project Report, Department of Psychology, Northwestern University, 1965.

The general outline of the intended study was thus as follows: For those invited into the Monteith entering class of Fall 1959, due to graduate June, 1963, and for the uninvited controls entering Wayne, institutional records such as grades, graduate record examinations, student activities, requests to have transcripts sent to graduate schools, registration for voting, etc., would be collected and compared. Since not all of those invited accepted, there would actually be three groups compared: Invited-Accepted, Invited-Refused, Uninvited Controls. Of these groups, for effects expected of Monteith, the first group should exceed the other two. However, even if this were found, it might not be evidence of effect. It might be that those who elected to accept Monteith were the more promising students, averaging better than the Invited-Refused group, and the unselected Uninvited Controls, for just this reason. If this were the explanation, then the mean of Invited-Accepted and Invited-Refused should equal the Uninvited Controls. This implies that if the Invited-Accepted were better than the Uninvited Controls, then the Invited-Refused should be poorer than the Uninvited Controls, a cue as to a selection effect. Randomization, of course, had taken place only at the level of invitation, not of exposure to Monteith. Thus the most conservative test of significance would be to compare the pooled Invited (Accepting plus Refusing) with the Uninvited Controls. If the Monteith experience produced a very big difference, it should show through in this pooled group, even though dampened or diluted by the inclusion of the Invited-Refused. It is possible that some more sensitive tests might be developed[4] but even if not, the study seemed well worth undertaking.

Problems in Establishing Comparison Groups

We knew when we undertook the study that the actual situation would be more complex than that outlined above. We knew, for example, that there were uninvited students who sought out and gained admission to Monteith (and who, in the most conservative statistical treatment, would have to be pooled with the Uninvited Controls, further dampening any effects). We knew that there were transfers both ways at various stages and frequently for reasons reflecting not at all on the adequacy of the programs transferred out of or into. But we were unprepared for the full complexity encountered when, late in the research, we achieved a detailed reconstruction of the invitational process. For it turned out that inviting and not inviting were end products of complex administrative procedures which followed principles of their own, most of them interfering in one way or another with the simple rationale of the scientific design outlined above.

After close examination of the original documents used in the administrative process, we have come to classify the problems under five different headings:[5]

---

[4] D. T. Campbell, "Invited therapy in an archival or institutional records setting: with comments on the problem of turndowns." Mimeographed project report, Department of Psychology, Northwestern University, 1965.

1. Not all curriculums were considered compatible with the Monteith program.
2. Quotas were pre-established for each of the curriculums deemed compatible.
3. Evening students could not take the Monteith (all day-time) classes.
4. "Extraordinary" students were not considered eligible.
5. Students whose files were kept for resolution of various unusual problems were not available for invitation.

We shall deal with each one of these points in turn--showing the nature and extent of the problem, and how we have dealt with it. In conclusion, we shall review some of the conflicts between the administrative and the scientific purpose, which were inherent in the recruitment strategy chosen, and suggest alternative strategies which might have eliminated some of those conflicts.

1. Compatible and incompatible curriculums

Besides the large college of Liberal Arts, Wayne State University comprises several other colleges: Education, Engineering, Medicine, Law, Business Administration, Pharmacy, Social Work, Nursing, etc. The college of Liberal Arts provides a general education for students who plan to graduate from, or to do graduate work in, the other colleges. The arrangements vary: Engineering admits freshmen directly, who take a number of courses in Liberal Arts; Education and Business Administration generally admit students who have completed two years in Liberal Arts; Law admits students who have completed three years in Liberal Arts; Social Work admits only graduate students.

---

[5] The greater number of students admitted in the Fall of 1959, and the presence of documents enabling us to follow the administrative procedure in the Fall better than that in the following Spring, have brought us to limit this detailed work to the Fall of 1959. The process can then be extended to the Spring.

When Monteith was created to experiment in general education it made agreements with the major colleges similar to the agreements these had with Liberal Arts. Monteith would thus prepare students destined to Education, Medicine, Law, Business Administration, Engineering. It would also accept a number of students interested in the "Liberal Arts" themselves. Now, these agreements were far from covering all the choices opened to a student seeking admission to Wayne. The Admission Form is reproduced below in a simplified version.[6]

- LIBERAL ARTS
  - General curriculum* (variety of majors, corresponding to Departments)
  - Curriculum in Science* (broad background)
  - Pre-professional Bus Ad* Dentistry Forestry Law* Medicine* Osteopathy Social Work Theology Education*
  - Special curriculum Art Chemistry Home Economics Humanities Journalism Mass Communication Music Physics
  - Professional curriculum Medical Technology Occupational Therapy
  - Certificate Secretarial (2 years)
- MONTEITH
- EDUCATION (a few special curriculums)
- ENGINEERING (several curriculums including Civil*, Electrical*)
- NURSING
- PHARMACY
- BUSINESS ADMINISTRATION (several curriculums)
- UNIVERSITY CURRICULUM Mortuary Science

---

[6] That year (1959) the form was altered in some important ways. Here we present the latest version.

Of these choices, only those followed with an asterisk were covered by agreements with Monteith. Thus, only the students applying for those curriculums were considered eligible to be invited.[7] Accordingly, seven different lists were kept in the Admissions office (general and science curriculums were not distinguished, but the two engineering curriculums were) on which was entered the name of each student admitted in the given curriculum, provided that he fulfilled other conditions (see below parts 3, 4, and 5). These lists record the dates of admission, invitation, acceptance or refusal, and other remarks jotted down by the person handling the invitations.

Our task here has been three-fold:

a) To make certain that an "invitation" was a genuine invitation: thus all civil engineering students had to come into Monteith (as a sub-experiment), they had no choice; they must not be considered "invited."

b) To distinguish between students who were invited to Monteith, and students who asked to be admitted to Monteith. These last are "self-starters," not to be considered as "invited."

c) To decide in cases when a student switches curriculum between application time and registration time whether he is to be considered "eligible" or not: usually, if one of the two curriculums is among those for which agreements exist, we have considered the student as eligible (e.g., a student changing from pre-Forestry to General, from electrical engineering to Physics, from secretarial studies to pre-Bus Ad).

2. The quotas and their consequences

In an effort, no doubt, to have a well-balanced student body, the planners of Monteith set quotas in their agreements with the other colleges. There would be room in Monteith for:

---

[7] This does not mean to say that students interested in Dentistry, or in Social Work, or in Humanities, or in Theology, could not fit in the Monteith program; some did. Only in an administrative sense did they fail to fit in the category of "eligible students." On the other hand, students interested in Nursing, or in Pharmacy, or in Art, did not have room in their program for the "general education plan" offered by Monteith.

70 students from Liberal Arts (general and science) curriculum
70 students from Pre-Ed
50 students from Pre-Med
50 students from Pre-Bus Ad
50 students from Pre-Law
50 students from Engineering (civil and electrical)

Now, these figures did not reflect the size of the demand for undergraduate education in these different areas--something which could have been ascertained from the admission figures in previous years. The difference between 70 and 50 gives only token recognition to the areas from which the largest contingents were to come. Thus, the "sampling" procedure could be expected to vary widely from one area to another.

In addition, invited students seem to have been variably disposed to accept the invitation. The propensity to accept varied according to the time of the year: some recruiting efforts were made in May, for instance, with some favorable results. Students closely scrutinized the details of Monteith's special understandings with various professional schools they were entering. (Thus Monteith had a much more favorable arrangement with the school of medicine than with Engineering.) Student response greatly affected the "sampling" procedure. Here is a rapid sketch of what happened for each curriculum.

a) Pre-education offered the best opportunity for a true sampling design. By the end of March 1959 more than 150 of the entering students eligible to Monteith had been admitted. One in two was being invited.

The 25 first invitations were sent to students on March 19. By the end of March, 4 had already answered yes, 6 no, and 15 had not replied. If the pattern set by the first returns were to continue, the quota was sure to be filled fast. Yet the rate of invitations was increased to about two in three.[8] As a result, the quota was filled very early: the last invitations were sent on May 20. Students kept being admitted to the pre-education curriculum until early September, but no one admitted after May 13 had a chance to get into Monteith.[9]

---

[8] Worse still, this decision was applied retroactively to the early entrants.

[9] Note the following exchange of letters:

> "May 22, 1959
>
> Dear Sir:
>
> Today a man came to Cooley High School to talk to us about Monteith College. After his vivid description and a pamphlet of the same was given to me, I felt very much desirous to know more about it. Could you please send me the Monteith College catalog to help me find the answers to many questions? Also, I want to specialize in teaching deaf children. The Cooley representative said there was a definite plan of study outlined for this field. I would be very pleased to receive this if at all possible. Any information in this line would be very much appreciated. Thank you.
>
> Sincerely,
>
> Jane Doe
> Detroit 27, Michigan
>
> P.S. I have just completed my autobiography for the application to Wayne University. I hope to enter college in September 1959."

The student was admitted on May 28; her application had been received April 29. A month later comes the following answer to her letter:

b) The General program was more of a problem. First of all, by the end of March, only 62 eligible students had been admitted. One in four had been invited. Of the first set of 6 students invited (on March 19), 3 had answered no, 3 had not answered yet, none had answered yes. Here again, the rate of invitations accelerated rapidly: by the middle of May and all through June, July, up to August 11, almost everyone was being invited.

c) In the Pre-med program, everybody was invited and practically everybody accepted, until the quota was considered filled. The last invitations were sent on July 1. All those admitted after the end of June were left out.

---

9 con.

> "June 26, 1959
>
> Dear Miss Doe:
>
> At this time we have no other printed material regarding the Montieth (sic) College other than that which you received from the representative who spoke at your high school. It is my understanding that the catalog is in the hands of the printer. However, we have not been advised as to the date of delivery. I would suggest that you write us about August 1 in regard to the catalog.
>
> I would like to add that at the present time we have ceased to admit students to the pre-education curriculum in Montieth (sic) College since the quota has been filled. However, if a change is made in the quota, we may be able to admit you should you express the desire.
>
> We are enclosing a Reply Sheet which you should send to us at the same time that you write for the Montieth (sic) catalog.
>
> Sincerely yours,
>
> Adam Smith
>
> Office of Admissions."

The "reply sheet" enclosed in the student's file is not one which was mailed to her but one which was automatically filled by the Admissions office, and marked "quota filled, please process admission 8-10-59"--i.e., process admission to Liberal Arts.

d) In the Pre-law, Pre-Bus Ad, and Electrical engineering programs, practically everyone eligible was invited, regardless of the date of admission. This came from the low number (compared to the quota) of Pre-law and Engineering entrants, and from the high proportion of refusals among the Pre-Bus Ad entrants.

Our task vis-à-vis quotas was to go back to the admission form of all students listed as having been admitted after the cut-off date in the Pre-ed, General, and Pre-med program. We established whether or not they ought to be considered "late applicant" and as such to be excluded from the "control group" of uninvited students. We found out that, in the Pre-ed program especially, many students listed as admitted in June and even in early July had applied in April (or, more exactly, before May 4, which is the last date of application of a student who was invited). We considered these students as eligible to be invited since the lateness of their admission was not due to them but to the Admissions office. The same procedure revealed that the same delay had excluded from being invited quite a few students in the General curriculum. In the Pre-med curriculum, however, the overwhelming majority of students admitted late had applied late, hence remained late applicants, not to be included among the control group of uninvited entrants.

3. Evening vs. daytime students

All Monteith courses were to be offered exclusively during the day, being based on a combination of lectures, given simultaneously to all students taking the course, and of small discussion groups. Students intending to take evening classes could not accept an invitation to Monteith, even had they wished to come; hence they should not have been invited. They should not have been placed on the lists of eligible admitted students at all. However, only the latest versions of admission forms allowed the student to make clear his intentions and possibilities regarding part-time vs. full-time and evening vs. daytime classes. Besides, students may have intended at first to study full-time, and found a daytime job later which made them change their plans. Thus the names of some future evening students were found on the lists of eligible students.

Our task here was to check carefully whether uninvited students were bona fide day students, or whether they had been by-passed during the invitation procedure because by then their intention of taking evening courses had been ascertained. The latter were not included among our "control group" of uninvited students.

4. The problem of "untypical" students

Now we switch to the consideration of uninvited students who were not put on the lists of eligible students even though they were eligible by all the preceding criteria. It looks to us as if the very definition which Monteith was attempting to give of itself, not an honors college but one for ordinary Wayne students, might have backfired. It looks as if by administrative translation "ordinary" had become "not extraordinary." Thus those foreigners (or adults) whose high school credentials were not easy to locate or evaluate (and hence who were given a "conditional" admission to Wayne) were not considered eligible to Monteith; nor were entrants who were allowed to take only a restricted number of hours until they proved that they were able to do work at college level.

We raise this question here largely to throw light on the discrepancies between the scientific definition of a "representative sample" and the administrative definition of "ordinary, typical cases." We tried to detect, among all "conditional" admissions of "atypical cases," which ones came closest to a regular program of studies. We found a good many such cases. We still did not include these in the control group, since not one single "conditional" entrant found his way into Monteith.

5. The time-consuming procedures

A great many cases of perfectly eligible students, however, are not found on the lists of eligible entrants, for one administrative reason or another. The files reveal a number of circumstances leading to the omissions. Most of them suggest that the student's file had to stay on someone's desk for a long time instead of being immediately forwarded to the person in charge of invitations.

A frequent case is as follows: the student has applied to Wayne earlier than in the Fall of 1959 but, for a variety of reasons, he does not register the semester for which he is admitted. Sometime in the Spring of 1959 he writes to reapply, is sent a letter of acceptance, and has his admission reinstated without ever being considered for an invitation to Monteith (even though the quota in his field is still open). Other causes of delay, hence of other miscellaneous omissions, are detectable: this student's name has been misprinted on his official record, or he wants to change his name; that student's high school record form is not the kind to which Wayne is accustomed, or his autobiography has been lost; this student's file has been kept in the Scholarship Office; that student has had an accident and hasn't taken an exam on time; this student has a "confidential file"; that student is changing curriculum; etc. etc.

Our task here was to ascertain which students' names had thus accidentally been omitted from the eligible lists, so as to place them again among those who "could have been invited to Monteith on time but were not"--hence among the "control group."

6. Conclusion

The administrators were trying to help launch a new college; they were not preparing for the scientific study of the impact of the college on its students. This basic orientation shows in many ways:

a) They looked in the directions most likely to provide them with a rather large contingent of students, neglecting minor curriculums (such as pre-theology, pre-social work, pre-dentistry) and thus limiting from the start the scope of any future comparison.

b) They set as their goal a specific number of students in each of the curriculums they planned to cover; thus they limited considerably the likelihood that the sampling procedure would be followed to the letter; for it is obvious that they were going to invite students until they fulfilled their quota, even if they had to invite every single one of those in a given curriculum (which is what happened in several instances as we have seen).

c) They acted with undue haste in some cases (pre-education), filling the quota too fast, and thus limiting drastically the size of what could have been a large control group, composed of students admitted at widely differing dates.

d) Finally, they tended to see the invitation of the proper students as just one step in the administrative process, which could be skipped if other steps had to be taken instead. There should have been a procedure whereby even cases held up by administrative necessity (change in name, delays in response from high school, etc.) would have been placed on the lists of students to be invited, and given a chance equal to that of other students.

In the light of the actual situation, the proper assembly of comparison groups presents many alternatives, each with disadvantages. If during one period, a random one in four is invited, and during a later period, one in two; and if the applications at different times differ in relevant ways, then even though Controls and Inviteds are randomly assigned, the pooled Invited and the pooled Controls will differ systematically. Early applicants tend to show foresight, and tend to choose Wayne as a first choice rather than as a reaction to rejection elsewhere. Due to the shifts in rates of invitation, early applicants would contribute more to the pooled Controls than would late applicants, most of whom were being invited. A comparable bias in pooled comparisons of data exists where different schools and curricula invite at different rates.

What we want to achieve is that every curriculum and every application period be equally represented in experimental and control groups, and that the assignment to one or the other be a matter of luck (the luck of order when every n'th one is being chosen, if not the luck of tables of random numbers, tosses of coins, or rolls of dice). We originally set up data analysis with the intent to employ all of the non-Invited as Controls, making the control group as large as possible. This was on the assumption that a constant invitation rate of one in three had been employed throughout the entire application season. Had this been so, we would have gained statistical precision by having the total two-thirds Uninvited in the Control. In the face of the continually shifting rates, and the differences in practices from school to school, it now appears that we must abandon any large scale pooling of Invited and Uninvited comparisons, and must go back to the original decision processes, designating a Control for an Invitee on a day-by-day, page-by-page paired basis, ending up with an equal number of matched Invitees and non-Invitees.

Thus for a school and period in which from applicant rosters one in four were being invited, some rule would be selected for indicating which of the Uninvited would go into the control group--possibly taking the next in order, or rolling a die, etc. When the ratio became one in two, all Uninvited would go into the control. When three out of four were being invited, then the experimental invited group would be reduced, only a random one of the three being used, to match the one Uninvited. Under such a rule, those schools and periods in which everyone was invited would contribute neither to experimental nor to control groups, since they were unable to contribute to the control groups. In a similar, but more obvious decision, periods in which no one was invited would contribute to neither group.

This is an interim report. At the present time, our recent reconstruction of the details of the invitational process has convinced us that our earlier tabulations are uninterpretable, but we have not as yet made new ones.

Problems of Access and Comparability in Archived Material

What is wanted are institutional records relevant to a comparison between Monteith students and the regular Wayne students. For this purpose, the unique advisor's and teacher's notes on students in Monteith College are useless--for there is nothing comparable in other Wayne divisions. Grade records in shared courses are probably relevant, although they are invariably Wayne rather than Monteith courses. But, in general, because so much of the Monteith grading is done within Monteith, and because Monteith's grading norms are unique, overall grade-point-averages are meaningless. Scores on the Graduate Record Examinations would be useful if equal proportions took it, but would be less interpretable if the

Monteith faculty pressured all of their students to take it while Wayne did not. For each potentially useful record it is necessary to examine the reasons which might make any differences irrelevant or misleading. While there are some general guide lines, basically this criticism must take place in terms of the specifics of any given record.

One argument sometimes advanced in favor of research with institutional records is cost. With interviews averaging perhaps $20.00 per respondent, it is easy to believe that institutional records represent a low cost method. No decision on this, or on the validity of records, should be made without an intimate understanding of institutions, their records, their record keepers, and the multiple cross-currents of fear, reciprocity, and chance which determine their quality and accessibility. In conveying this reality, it has seemed to us useful to provide extensive excerpts from an activities log kept during the course of our data collection.

June 5, 1963 - Had lunch with A regarding (1) transcripts. He said it was a lot but they'd do it if cleared through B. He also suggested when we got list we check it against Spring Alphabetical roster and identify for benefit of his office which students were active and inactive. We agreed transcripts for actives would show grades through Winter Quarter; (2) who would have list of FTIAC's (first time in any college) we wanted. He suggested C because Unit Z prepares a list of freshmen admitted and registered. After lunch back in office, I phoned C who said he didn't have such a list. He referred me to D. D said they had prepared such a list in Unit Y for E and he'd check with her. Later he phoned back to say E had the list and we could borrow it. D also gave me names of offices which received copies of Alpha roster. I phoned A, Unit X and F's office telling each we would want it to keep--no luck on this basis, but all would loan it. Had interview with G--agreed on $2.00 per hour pay rate but told him (1) that I wanted to postpone his starting to conserve money for a possible second person and also clarified fact I didn't know how long job would take.

June 13, 1963 - Asked H to clear research angle with (1) I who was to pave way with J and F; and (2) B who was to clear with E, K, L, M, N, and O. Also called F who said she'd give me her Alpha roster and would drop it by my office 4:00 p.m. Monday.

June 17, 1963 (Monday) - Had 10:00 a.m. conference with E and Q. Q said he'd have to talk with M or L regarding transcripts. Q also said he'd need clearance from Unit X if transcripts of their students were to be made. E had her girl go make two thermofax copies of FTIAC list. I helped girl. Went to lunch with A. He said he expected grades in his office on 21st but if I could get list to him promptly and if B cleared he could have all transcripts done by then. However, list would have to reach him soon. It was agreed that neither of us could mention this and had to wait until B moved.

After lunch I had H phone R for Unit X clearance with B. R delegated S to do this. I started G to work when F brought Alpha roster. He began checking "actives" on E's list that afternoon. Telephoned B to ask if Unit X clearance had reached him--(this also was to give him a prod regarding transcripts). Secretary said he was in conference but she'd interrupt him if it was urgent. I said "No" and we agreed I would telephone him at 8:30 a.m. on the 18th.

June 18, 1963 (Tuesday) - Phoned B who said he hadn't had time to discuss matter with L but would do so when he had time.

After lunch returned to office to find there was a snarled up class list that would require my going to Unit W. After taking care of that I checked with A as to whether or not he'd received clearance from B. He had not! Was lucky and ran into B to ask him about it and he said we should go see L right then because he hadn't had time to do so. Spent considerable time with the two of them. She was discouraging, claiming that it was too close to end of quarter and that she thought it couldn't be done during regular office hours. He was concerned because of his budget, which would not provide for overtime help. They told me they'd have to work out a plan and telephone me. I left, disgruntled because B hadn't checked with N. Had parking violation ticket on my car. Returned to office. B phoned to say he and L had decided they'd have two girls start Saturday on the job and they would work succeeding Saturdays until job was done. I was extremely upset because I had reason to know this would take at least five weeks which would delay our progress but I couldn't express my disappointment in their decision.

June 19, 1963 (Wednesday) - About 9:15 a.m. L phoned saying, "How soon can you get that list up here? N says they could start on it now and work until grade slips arrive." I said I'd drive it up immediately. G was recording actives on his duplicate copy of the Fall 1959's. We got that list to A around 10:30 a.m. After lunch G and I went back up with Spring 1960 list and kept a prearranged appointment with J. He told us he was leaving on the 18th for two months in Europe. As an example of data he might not wish us to have I mentioned the Confidential Parent Income form required as part of Student Financial Aids application and Counselor's Comments. He said he saw no reason why the data shouldn't be available to us because "...If you couldn't trust researchers who could you trust?" He gave us carte blanche with all data and then escorted us to U. U gave us copies of Financial Aids forms and showed us how their records were kept. J then took us to V. Her data were kept in summary form in raw scores on index cards and on typed summary lists by testing dates. We agreed latter would be best to use because of ease of thermofaxing. G made arrangements to do that Friday morning. She gave us copies of her test profile forms and index cards. J then took us to W. Their data were on summary cards and four page 8 1/2 x 11" records. This concerned us because of a duplication problem (ball point pen ink won't thermofax). W gave us copies of their test record forms. We next met with X in Unit V. He showed us how his job applicant records were filed on cards; however, his files were not complete back to September 1959 -- it had been necessary to throw away cards that old because of space reasons. His cards dated back to July 1960. The cards were small and again could present us with a problem regarding copying of data.

G and I returned to my office and I agreed that Y could be hired to type for the project at $2.00 an hour. It had seemed to us that tabulating sheets would be needed to check and record data on the 1633 students who comprised the sample. Y, G and I agreed that stencils should be typed and run off using the name and I.D. numbers. While there were many ways it could be done we chose to type from E's original list, across the long way of the paper and retaining the grouping according to original college.

June 21, 1963 (Friday) - A phoned to say they had copies of all of the transcripts excepting for "about a dozen which had not been located -- probably due to name changes." He was very pleased and I complimented him on their efficiency. I had G go up for the transcripts because we had implied there was urgency to getting them and I didn't want to dispel that notion.

June 24, 1963 (Monday) - I telephoned J regarding his written approval on Confidential Income form of parents. I couldn't believe his answer!!

"You will have to get the approval of Z for that," he said, "because there has already been one case involving that data pertaining to one student. Maybe if you can convince him that it will lose identity when grouped." "What about the Counselor's Comments?" I asked. J replied, "You'd probably have to get approval from AA." I couldn't resist saying, "But didn't you say it would be alright and 'If you can't trust researchers, who can you trust?'" He said, "Yes, but since there has been this one case involving Z you'd better check with him now."

I telephoned V and apologized for G missing his appointment with her last Friday. Even though I explained the reason she was slightly difficult and said she had such a busy schedule she didn't know when she could see him. I suggested that perhaps she could leave the lists to be thermofaxed with a secretary but she said they'd be too busy to be bothered. She's certainly going to get even with G for standing her up! Finally got her to schedule an appointment with G for 9:30 a.m., July 3rd. She's not that busy but she's made her point. I told G he'd better be there even if he has a broken leg.

G worked most of the day collating tab sheets which Y has begun to run off.

June 26, 1963 (Wednesday) - Took BB to lunch. BB is with I.B.M. After lunch BB accompanied me to the Unit U and introduced me to CC. CC gave me a copy of two different lists. One was all of the data currently on tape regarding WSU students and the other was a much more elaborate listing of all of the items they hope to eventually have on tape.

G finally finished collating the tab lists -- both he and Y have done a big job.

I told G that I was getting alarmed because it was nearing the end of June and he hadn't started yet to obtain data in a single office. I have the feeling that he is not looking forward to this part of the job.

Telephoned E's secretary and asked her if she'd thermofax two more of the original lists when she had time. She agreed to do it soon and mail them to me.

June 27, 1963 (Thursday) - I met with G first thing this morning to discuss what office he was to start with. I telephoned U who said he was too busy to start G but would see him Friday. On the assumption that G could pick up Unit V any time (I know the secretary in that office quite well) I telephoned Unit T and talked with W who is apparently in charge while DD is away. We worked out a means by which G could begin by checking which students had tests other than Freshman or Entrance or Placement exams, these being in V's domain. G went up there and telephoned me just before 5:00 p.m. to say he'd almost finished this job.

July 1, 1963 (Monday) - G and I met for about an hour this morning. He had checked through most of the roster on Thursday for those taking special tests resulting from students' voluntary consultations in Counseling. On Friday he had spent most of his day in Unit S with U but had left there early enough so as to go back to Unit T and finish checking names. W had told G he'd like to know his results so G had come up with the figures, which were impressive. I shaped up the figures into legible chart form and they indicated that 60 or 18.86% of the Monteith students, 231 or 18.33% of the Liberal Arts students and 11 or 33.33% of the Engineers had taken advantage of this University service.

At 1:30 I kept an appointment at Unit Z with CC and C, to clear access to the Fall 1959 and Spring 1960 records. We concluded that the cover and inner left-hand page of the double-fold application form were not of significance for our purpose but that the right-hand inner page and the back were. I was told that these along with the high school transcript could be copied or reproduced but that the student's autobiography and the high school principal's and counselor's recommendations could not. They pledged their willingness to cooperate and told me that the files were alphabetized by year. I agreed to introduce them to G when the time occurred for him to begin in their office.

I called X about G's working there. We had cleared everything before so I was expecting him to simply say, "Sure, send him up," but instead he said, "Have him come up and we'll see what we can do for him." I regarded this as a little peculiar because it could have meant there were going to be some restrictions of some kind.

But when I returned to my office G telephoned to say that when he'd arrived at the Unit V, X had agreed to give him all of their placement applications from July 1960 to June 1962 and G wondered if I wanted him to bring them all or just those for the names on our roster. I told him to bring them all back and he could sort them out back at our office. Bless X's heart -- I had certainly misread his remark. This must have been what he planned to do for G. It will save G much time and work for he'll only have to copy or reproduce the Active cards from 7/62 to 6/63. G brought back all the cards with him.

July 2, 1963 (Tuesday) - I was so pleased with what X did and I thought about it last night at home. If he could give up old records it would do no harm in at least raising the possibility with other offices. I telephoned CC in Unit Z as soon as I arrived in the office. CC said he'd check with C if I'd hold the phone. He came back on the line and said C said they were supposed to keep them until October 1963. I asked him if it really made that much difference, July or October, and he said he thought not. So it was agreed that G could have the Fall 1959 Admission records if he would cut off the principal and counselor recommendations and pull out the autobiographies. They could not, however, give up the Spring '60 records. I was so elated because the Fall records were over two-thirds of the total number of our sample.

When T arrived I told her of our good fortune but explained that we still had a problem. If G had to copy material by hand from forms he would benefit from any shortcuts we could suggest. Did he have to have all the data in its complete form? T spent about forty-five minutes with me suggesting items which could be deleted or copied in an abbreviated code form. Even so, it seemed like a large task for G to do by hand yet our thermofax machine would not reproduce ball point pen or colored inks. T and I had both seen an advertisement for a new SCM duplicating machine which would reproduce colors. T telephoned J to ask if she could arrange for a demonstration of the new machine. T departed and J called later to say that the machine would be demonstrated in her office at 2:00 p.m. Wednesday.

G spent most of the day separating the placement application cards of the sample from the total group.

July 3, 1963 (Wednesday) - G kept his 9:30 a.m. appointment with V and thermofaxed the Freshman and Entrance exam test score lists of all students admitted Fall 1959 and Spring 1960.

At two o'clock G and I joined T, EE and FF over in J's office for the demonstration. The salesman was embarrassed because he'd forgotten to bring some part and had to return to his office to get it. We reassembled at 2:45. The machine was a big monster, cost around $1,000. The shortest leasing period was one year at $80.00 per month or five years at $17.00 per month. J was to leave on vacation and left the matter with T and me to consider and decide.

July 8 - 12, 1963 - G used the early part of the week completing his check on the transcripts. He found that four or five were missing. I called Unit R and told GG (A's assistant). He had them delivered by messenger within an hour. Originally there had been seven other transcripts which they could not find. Little by little they had found one, then another, all were name changes. Finally they quit and admitted they had exhausted their search for the last three. I spent an hour sleuthing and was successful. I telephoned Unit Z and had a clerk give me the addresses and parents' full name. By means of the city telephone directory I called the parents in two cases and a cousin in the third case and learned the married names. Something had just gotten fouled up in the office's cross-reference system. I kidded them and they mailed me the last missing ones.

G discovered many discrepancies in his original list of the sample FTIAC's and he came to me for help. There were cases of students who had taken college work at Wayne as non-matriculated (a category which was dispensed with when the Wayne-U. of M. Adult Education Division was formed) or Trial Program students before Fall 1959. I wasn't positive on the University policy of the FTIAC's definition in this regard. After several phone calls I ended up with B who said to use the trial program students who were put in matriculated status either Fall 1959 or Spring 1960, but not the non-matrics. (He explained why not.) By deleting the non-matrics, duplicate names on lists, etc., the sample group now consists of 1612, according to G.

At my suggestion he figured out and had Y cut and run a stencil to cover all the corrections to the original roster.

On Friday morning I received a telephone call from the salesman of the Xerox Corporation. He said he was on campus and A had told him we had a duplicating problem. We arranged for a demonstration at 4:00 p.m. that afternoon. By that time I had reached T and she had FF go with G and me. The salesman took us to HH's office where they have one. It was a fabulous machine and would have been simply great. It sells for $29,500! But they would rent on a month-to-month basis at $25.00 per month plus 3 1/2 cents per copy for a monthly minimum of 2,000 copies. I thought we had found the Ideal but then the salesman said it takes about six weeks to get one -- probably would be mid-August so that killed that!

July 15, 1963 (Monday) - At 9:30 G and I went to meet a salesman for the Verifax Copier machine. It was a compact, versatile machine which, except for the folders in Financial Aids, would have been perfect for our use. Even the purchase price of $549.00 was reasonable. However, the machine has just been introduced and would not be available until mid-August at the earliest. When the salesman realized the urgency of our need he volunteered to loan us (the inference was for free) a machine which would do the job. We would have to buy the paper from him. He agreed to meet us at 10:00 Tuesday morning in Unit S to see the kinds of records we wanted to copy.

After lunch I telephoned U and asked him if he would be available at that time Tuesday morning. He would not be because he would be tied up with auditors. I said I was just checking because we were meeting the Verifax salesman in his department at that time and I had thought it might ease his mind if the salesman assured him that reproducing would not damage his records. I also told him that we could do him a favor by partially clearing out his inactive records if he would like, and that Units Y and Z had offered to let us have the old student records in our sample. He hemmed and hawed to the effect that maybe we could have some but records in the inactive file weren't necessarily inactive in the usual sense. He would be willing, if G pulled all folders, to then go through them and tell G which ones we could have -- there was no special criterion he could explain to G so that G could do this himself.

July 16, 1963 (Tuesday) - G and I met the Verifax salesman at Unit S, and were in and out of that office for over an hour, and there wasn't a soul with U the entire time. Some story about the auditors! The salesman looked over the materials which were to be reproduced there and we decided that it would require 12" x 18" paper to do the folders. This size is expensive (18 cents) but worth it in another way because both inner sides of the folder can be run in one operation. The miscellaneous material will be run on 8 1/2" x 11" paper and a number of small items can be put on the same page. From this office we went to Unit Z where we met together with C. He showed us where the Spring 1960 records were kept and indicated a location where G would be able to work with the machine. Since we would want to reproduce the right-hand inner side of the folded application form and then the back side we realized that two 8 1/2" x 11" pages would be required, and a third one to duplicate the students' high school record approximately 1300 pages in all. G returned to Unit S and the salesman and I went to Unit O to see II. He advised us on how to proceed in order to expedite the order so that the salesman could bring the paper and the machine with him 10:00 Friday morning. I also got the salesman, in front of II, to promise to credit us with any unused paper in which the 100-page package was unbroken and also to give us fast service if we needed to re-order more.

Earlier in the afternoon I had been thinking about the salesman's statement that many small pieces could be reproduced on a single sheet of photographic paper. I had previously learned that when a student went to Unit T and special tests were taken that his test scores were recorded by V on a small 3 x 5 card and also on a large 11" x 17" double-folded profile. I had thought it would be easier, quicker and more economical to reproduce, e.g., six of these, at one time on a page rather than one profile at a time on the large paper. I telephoned V to ask if we could have access to "her" 3 x 5 cards rather than the profiles which the Counselors had in their files.

(The reader may recall that J had cleared with V for the Freshmen or Entrance exam scores but with W for the Counseling Profiles.) V said that under no conditions could we use her cards without written permission from each student whose scores were involved. She went on to explain that when each student voluntarily went to Counseling and took tests at the recommendation of the Counselor he was informed that the test results would remain confidential between him and the Counselor. She explained that tests which the University required of the student were different in this respect and could be made available to faculty. She said this had always been the policy of their department and that if we did use the profiles in the Counseling files DD would be extremely upset when he returned from vacation. I explained that JJ had cleared the matter with KK who in turn had cleared it with J who had cleared with W. V said that she dounted if J even knew what went on in their department and that he never came near them. I then explained that I hoped she wouldn't take it personally but that since we had previously cleared with, and been assured by, those whom we felt had the authority to give clearance I was obligated to pursue the matter further. She indicated she quite understood. I then phoned W and explained what happened.

T and I discussed two other matters. She asked how up-to-date the student activities records were in F's office. I told T that F's secretary had indicated that they were recorded annually each summer. T suggested that we then schedule that job for later in the hope that when we did do them they would be up-to-date. She also asked if I could predict a completion date for the data gathering and I said I could do so better when we observed how we fared on reproduction.

July 17, 1963 (Wednesday) - I placed a call for the Verifax salesman early in the morning regarding the permanent non-fade quality of records reproduced on his Dri-Stat paper. He assured me that it would make permanent non-fade copies. I informed him that II in Unit Q would have the order by noon. G drew up an estimate (on the low side) of the paper we would need -- 2,000 sets of 8 1/2" x 11" -- $160.00, and 700 sets of 12" x 18" -- $126.00, which along with the fluid totalled $287.20. LL typed the order, MM signed it and G took it to Unit Q. I checked with II by phone later to make sure he had received it.

W called me in the afternoon and asked if I hadn't mentioned an official letter which verified the sponsorship of our data collecting. I said yes, I'd send him a copy. He replied that would be fine, he just wanted to have it to show DD when he returned. I asked if the matter was cleared up and he said, "Not yet."

July 18, 1963 (Thursday) - I have been remiss in not mentioning that G habitually clears his previous day's progress with me. We also try to work out solutions to problems he has encountered. One of the points I stressed with G was the importance of Y's working with him because much of what he's now doing would go three times faster if two people were doing it. U has agreed to go through the folders after G has pulled all of them and decide which ones G could keep. G is confounded (and I don't blame him) because he has found a number of instances where there has been a folder for one student in both the Active and Inactive file drawers. In one instance he found three folders on one person. His difficulties have been confounded by name changes of students, not only resulting from women students who have married but an occasional male student who has abbreviated or Anglo-Saxonized his name. This necessitates G's checking under both the former and the present name because on the transcripts the old name was eradicated and the new name was inserted with no date as to when the change occurred.

Before noon G asked U's permission to bring Y back with him when he returned to Unit S after lunch. G and I had discussed how he might approach U on the basis that he would be able to get out of their way sooner if he had help. G told me that U agreed reluctantly and commented that he would like to keep the number of persons who perused the files to a minimum.

W telephoned me close to five o'clock to tell me that everything had been cleared -- he had talked with AA.

July 19, 1963 (Friday) - G went up to Unit S early this morning and I went there so as to be present when the salesman brought the machine and paper which he promised to do at 10:00 a.m. The salesman was late and while waiting I observed U, whom G had been able to start going through the Inactive folders. G seems to have established good rapport with the two office girls and with U too. Occasionally the latter would say, "Here is a folder you can't have," and G would murmur, "Okay." I said, "U, if there are such instances could you give us a general reason why not so that G can note it?" He said, "Usually they are cases where mental hygiene has failed to approve them or there is something else regarding their record of a confidential nature. In any case it has prevented their being approved for financial aid."

The salesman appeared at 10:30 a.m. and said he'd been downstairs trying to gain access to a freight elevator. I was surprised because U dispatched a fellow who's helping in the office this summer to go with the salesman and help him. Soon the two reappeared with a freight truck. The salesman apologized that he hadn't been able to bring the large-size paper with him -- he'd bring it next Monday or Tuesday. This was disappointing because it would mean that while G could start doing the contents of the folders, he'd have to backtrack and do the inner sides of the folders later.

I was sorry to see that two operations were required for a positive print and that they were quite wet. The salesman said that he'd bring a drying rack when he returned with the larger paper and that they really dried quite quickly. He was right, and the copies did not smear when wet. I left when I was assured that G was catching on to the machine's operation.

I received a carbon memo from A to BB on the estimated cost for producing transcripts for our use. A carbon had also gone to B. The cost was $110.55. I doubt if we'll be asked to pay and I think B wants to use that as an example of why his division needs a larger budget. Originally I had quite clearly stated to him that we wouldn't incur a financial obligation unless we knew what the cost would be and he had hedged. I decided to bluff it through with him (BB is on vacation) so telephoned and asked if he thought we owed him $110.55. Again he hedged and said he'd talk to L after she got back from her vacation.

July 22, 1963 (Monday) - G and I met again this morning before he went to Unit S. There is far more in these folders of students than he had anticipated both in terms of our time and paper estimate. Since it appears impossible to establish a criterion for determining which documents will and will not prove of subsequent value in the research phase of the study we have arbitrarily agreed that G duplicate everything in the record which we've had clearance to have.

G returned to my office at 3:00 p.m. and brought copies of those records which he had photographed. He showed me cases which included 10-20 pieces of paper and told me it had taken him and Y all afternoon to do the contents of 27 folders. We tried to estimate a completion date for Unit S and the other offices he had to do and we were into the month of September.

July 23, 1963 - This morning I figured out as best I could what money we had to work with and then called T for permission to hire another person. She gave her approval and at 1:30 p.m. I interviewed OO. I gave OO a brief orientation on the project and she accepted the job at $1.30 per hour. I drove her up to Unit W, and had her wait for me in the Unit Z Office while I went to get G in Unit S. He had run into operational difficulties with the machine. It functions but the copies are not as clear as previously and the fluid level on the machine is out of kilter and requires more fluid than it should. Also, he does not yet have the large paper! I telephoned the company to leave a message for the salesman, and the secretary asked to talk with G to see if she could give any instructions which might help. I left G talking with her and asked Y to have G come to Unit Z when he could.

I rejoined OO and introduced her to C who took us to the files for 1959 entrants. I requested OO to start pulling the folders which we can keep, providing certain materials are removed and cut out with scissors. A room with a table has been made available to her and as C was going through the folder showing her which things we could take and not take we were joined by G.

At 3:30 p.m. I kept an appointment with PP, Director of Unit P. I first gave him a copy of the letter of authorization in order to provide verification of the purpose of our role in the project and then spelled out our specific needs with regard to data collection. We first discussed an enclosure used in the University application in Fall '59-Spring '60 relating to the entrance physical. PP pointed out the printed word "Confidential" which appears on the upper left corner of the back page. It is this page which Unit P eventually receives, whereas the first page is detached and retained by the student. Because of the "Confidential" notation, PP felt that it would be a violation of this assurance to make the form available to us.

Two other forms also are kept in each student's folder. One is the lab report record which is part of the admissions physical and one is used to maintain a cumulative record of a student's visit to Unit P. PP indicated that unfortunately these were also regarded as "...privileged communication between the physician and the student." I told PP that I'd heard it might be possible for us to determine how many visits the students in our sample had made to Unit P. He said that if we were to do checking it would mean opening the students' folders and he could not permit this. "Yes, the dates of a student's visits probably would be all right, but what the student came to be treated for and who treated him would not," claimed PP. I pursued the matter further. If one of his staff were to look into the folders and read the number of visits to one of our recorders, would this be permissible? He said his records office was already overworked and understaffed and he was afraid not. "Besides," he continued, "of what significance would that be -- a student making only one visit could have been so seriously ill he had to be referred to a private hospital [and we could not be told why] while a student could visit a number of times about something minor." I asked him about visits to Unit O and he said that they were even more confidential but that I could discuss it with QQ, Director of Unit O. I thanked PP and went to QQ's office. I followed the same introductory routine with him as I did with PP. QQ gave me a copy of a form which students complete when seeking their counseling services. Those forms are later filed in a student's folder along with an ordinary stenographer's spiral notebook in which the therapist keeps notes. In answer to my request for data of any kind pertaining to these cases he replied, "I'm sorry, but absolutely no, -- we don't even give information about this to the F.B.I." He then cited a few historical instances where their office had revealed information to supposedly reputable sources only to have this trust broken. "Never again," he said.

G returned to my office at 5:00 p.m. The Verifax salesman had telephoned him and said he'd try to get to G tomorrow and if he could would bring a repair man with him. G continued using the machine but it was not wholly satisfactory. G had checked with OO before he left and felt it was a good decision to hire her. I almost forgot to mention my pleasure at seeing G eating lunch with RR, the young man who is assisting U this summer. Don't know if G did it consciously or not but I complimented him on maintaining good relationships with the personnel in the office.

July 24, 1963 (Wednesday) - G talked briefly with me this morning. As yet, the large photographic paper has not arrived. This is beginning to be increasingly inconvenient. It may be remembered that most students applying for financial aid had an application form which was printed on the inner two sides of a letter-size manila. It had been decided that in order to minimize damage to the folders and to save time we would use 12" x 18" paper and photograph both inner sides in one operation. In the meantime G has been duplicating the contents but must keep out the folder to do later. It would be easier if they could have been re-filed as they were finished. I telephoned the salesman who said that the manufacturer had shipped it and we'd have it by the first of next week.

July 25, 1963 (Thursday) - I should explain that because OO comes to work by a bus that picks her up and drops her off at Unit W she does not check with me. She is working a full day five days a week pulling the Fall 1959 Admission folders and deleting the material which must be cut out or taken out before we can keep. G works in the same building but on another floor.

G came back close to 5:00 p.m. loaded down with folders which OO had pulled thus far and with stacks of pages which he had photographed. He dumped all of this on the top of my desk and with a big, proud grin said, "D'ya want some data?" He seemed happy and pleased and I think is liking the work better. He reported on OO's progress and told how fast he and Y were working this afternoon. "We really got an efficient production line going this afternoon."

July 29, 1963 (Monday) - G didn't check with me until evening. He still does not have the large paper and is running low on the 8 1/2" x 11" paper. I told him I'd check with the salesman first thing in the morning. G thought he had enough 8 1/2" x 11" paper to last through the 30th.

July 30, 1963 (Tuesday) - I telephoned the Verifax salesman at 9 a.m. and explained the urgency of G's getting the large paper by Wednesday morning. Besides holding up G, U has told G to have all of the folders back in the files by Friday a.m., because their office needs them (and they really do!). The salesman agreed that if the paper which had been ordered had not arrived he would bring another type of paper that size along with a different kind of developer fluid. I also told the salesman I was processing a supplementary order for 2,000 sheets of 8 1/2" x 11" Dri-Stat paper at $160.00 and 6 quarts of developer at $7.20.

I went over to LL's office only to learn she was on vacation but TT typed up the purchase requisition. Couldn't find either T or MM to sign it so left it with Y for JJ to sign. I told her when it was signed to give it to G when she saw him at lunch and have him leave it with Unit Q when he returned to Unit W.

About 2:45 p.m. I received a telephone call from G. "I'm all out of regular-size paper," he said, "and I've got to have more right away." I asked if he'd taken the order up to Unit Q and he said Y had mentioned it but said that JJ wouldn't be able to sign it until he'd returned from lunch. G was obviously quite impatient with me because I couldn't promise him some paper within 15 or 20 minutes. I said I'd try and do my best. (I was annoyed at him for waiting until he was out of paper before he called.) Anyhow, in the end it all worked out like a miracle. I telephoned II who said if we were going to do it wrong we might as well go all the way, and he gave me a purchase order number over the phone and I called the salesman's secretary and told her we were in a jam and needed a paper order delivered as quickly as possible. I gave her the order and the purchase order number and she said they would have it at the W.S.U. Receiving Room within a half hour. I telephoned G and told him to meet me out in front of Unit W and I drove up with the purchase order for him. I wanted him to deliver it personally to II. I drove him around to Receiving and told them that as soon as the order arrived they were to deliver it to Unit S. They agreed, and G left me to take the order to II. About 25 minutes after I returned to the office G telephoned to say that the order had arrived.

July 31, 1963 (Wednesday) - I didn't see G when I arrived this morning but Y told me he had gone directly to Unit S. I was in conference with JJ when at 9:45 a.m. an "urgent" telephone call came through for me. It was G who said that the salesman hadn't arrived with the large paper. I told him to call me back at 10:30 if the paper had not arrived. I learned later that the long-awaited Dri-Stat paper which we wanted had come and it had not been necessary for another kind to be substituted.

G and Y came back to our building after 5:00 p.m. and worked, folding the large paper so as to use it as a folder to hold the other forms and correspondence for each student. As they did so, they tried to put the data in the sequence in which the students' names appeared on the roster, so as to facilitate the subsequent job of the data tabulator.

August 1, 1963 (Thursday) - I met with G as soon as I arrived at the office this morning. OO has now pulled all of the Fall, 1959 Admissions folders, clipped and removed the portions we cannot have, and G has brought them all back to our building. He now has OO pulling the Spring folders and arranging the contents in an order which will allow him to photograph it more quickly. It was G's feeling that he and Y could be finished by tonight in Unit S. While I was talking to him, T telephoned about another matter and I asked her if she'd like to see the kind of data which G had obtained. She said she would, so G took a sample of the Unit S records to her office. T also asked me about the feasibility of UU checking the sample names with the City Clerk's Office as to whether or not each student was a registered voter. This came as an unexpected surprise so I did not answer her directly except to say that if it had to be done I'd be in favor of having someone other than him do it. She said I should give it further thought and she'd discuss it with me tomorrow.

August 2, 1963 (Friday) - G was waiting for me when I arrived in the office this morning and was happy to say that they had finished yesterday in Unit S and had begun pulling testing profiles. He said he'd like to keep the machine where it was in Unit S and bring in the testing profiles to reproduce from Unit T which was next door. I telephoned W to ask if it would be agreeable for G to work at their files today and he said that it would. He also approved of the profiles being taken into the adjoining office.

I talked with T briefly about the check on voting registration and advised against UU doing the work. We agreed to wait until the campus data had been collected after which time perhaps OO could do it.

When G returned at 5:00 p.m. he told me that Unit S had needed the space where he had been working for sorting out the files of scholarship applicants this Fall Quarter. He was then able to arrange adequate working space in Unit T through the cooperation of W.

G made one disappointing discovery, however: from the 3 x 5 cards which had earlier provided his record on the presumptive list of profiles there would be, he learned that quite a number of students had never followed through

6

and taken the battery of tests which had been recommended by the counselor. This would seem to suggest two possibilities: either the interview had been adequately satisfying to these students, so that they did not feel the need of the tests, or the interview had not been satisfactory and these students therefore terminated their contact. While I don't know if all counselors follow the same procedures during their initial interviews, it should be pointed out that failure of some students to return for reasons of dissatisfaction is not necessarily an indication that the counselors are at fault. Too many students are naive about tests and anticipate that results will provide a panacea to their problems. I have always respected the practice of this department in the way the personnel straightforwardly explain the limitations of tests to students. It is conceivable that this revelation would discourage some students from investing further time and effort.

August 5, 1963 (Monday) - I did not see G this morning but met with him around noon. I expressed my anxiety about the working schedule and tried to get him to estimate how long it would require to finish reproducing the testing profiles and the students' records from the current files in Unit V. He was hesitant to be too definitive. I explained that because I would be taking off vacation days until the 13th, I was concerned about G's leaving the project on the 16th and wanted as much completed by then as possible. Y entered the office while we were talking and I suggested that G take OO away from Unit Z to pull the Unit V cards in preparation for his arrival there. He did not readily agree but Y said she thought it was a good idea so G agreed to do it. OO started doing that in the afternoon and just in time because G completed the test profiles and moved the equipment to Unit V.

August 13, 1963 (Tuesday) - Talked with G in the office this morning. There were about fifty-five of the Fall '59 admitted students whose records had been found by OO in the 1960 files. One of the supervisors in Unit Z had told OO we could not be given those records because they were "Conditionals." This would take considerably more time out of an already tight schedule and G asked me if I could intercede. I telephoned the supervisor in question for an explanation of her reasons. Her basic reasoning was valid but there was one contingency which she was not taking into consideration. I called this to her attention and she admitted that those students who had subsequently achieved matriculated status perhaps could be regarded as Fall 1959 admissions. She suggested I get clearance from VV (Director of Unit Z) who has now returned from vacation. I telephoned him and explained B's (his boss's) definition of a FTIAC and how consequently I felt that technically this entitled us to the actual records of those who were now matriculated. He agreed with my logic, said he would tell WW, but that it would "...cost me a fifth." I told him I couldn't even afford one for myself but that I'd buy coffee sometime. His approval will save us much time because most of these cases were now matriculated.

G also explained to me how he had "paired" students' job application file cards which had data recorded both front and back. He had reproduced two front sides on one 8 1/2 x 11 sheet and two back sides on a second sheet. This showed good thinking on his part. He also said that the cards were of a colored stock and didn't reproduce as well as white.

I was then very surprised to receive a telephone call from L regarding A's memo pertaining to the cost of reproducing the transcripts. She asked me what arrangements we'd agreed upon about our paying the bill. I said I had not committed us to anything because B had been unable to provide me with an estimate. L said, "Well, I know you have budget problems and so do we. Would you be willing to split the bill?" I said, "From your standpoint would you regard that as satisfactory and fair?" She said that she would. I got her account number and agreed to check it with T. If T approved, I'd see that $56.00 was transferred into her account.

I telephoned T who was at home and just leaving for a trip for the remainder of the week. I explained to her that the metal storage cabinet in the "Chapel" of our building was almost empty and I would like G to store the data he has collected in there and lock it. She gave her permission to do this. (There was a key in the lock so I asked G to get a duplicate key made and said I would pay him later.) I told her about L's call and she approved of my having XX transfer the funds from the Program Study account.

August 14, 1963 (Wednesday) - I met with G again this morning and outlined all the things I wanted him to do before he finished Friday:

(1) Complete his log.

(2) Instruct OO as to all operational phases of the reproduction machine.

(3) Make certain that all data were grouped properly with a check-off student roster on the top of each stack to facilitate subsequent identification, i.e., knowing that all data found were in the same order as the students' names were checked off on the roster.

(4) Give me a summary of number of subjects for which data were found in each of the University areas covered.

(5) Total up OO's, Y's, and his pay to date.

(6) Continue getting as much work done as quickly as possible.

He told me that there were 250 sheets of 12 x 18" paper left in unbroken packages (the salesman had agreed to take back paper in this case) but that the 8 1/2 x 11" paper stock and the supply of developing fluid was low. The paper to be returned was worth $45.00 so I called the company and told them I wanted to swap the paper for 500 sheets of 8 1/2 x 11" paper ($40.00) and six quarts of developer ($5.25) and to tell the salesman I'd give him a quarter for the difference the next time I saw him. We made arrangements for them to bypass the Receiving Room and deliver the supplies to Unit Z before 11:00 a.m. Friday and G would give them the large paper in return.

August 15, 1963 (Thursday) - G reported that he had timed OO on the machine and because of the number of pieces to be reproduced in each folder and the slowness of the process, she was averaging 13 to 15 folders an hour. He said there were about 275 left to do. He reminded me that refiling should also be considered in estimating job completion. He left the number of hours each of the three of them would be reporting for the pay requisition of August 23. His time and probably Y's would actually not be beyond the 16th. I told G I'd like to hire YY and G thought it would be a good idea. He said that the work went better with two people to operate and feed the machine and he thought a man could clean the machine better than a woman.

August 16, 1963 (Friday) - At noon G gave me his log, the summaries of the data gathered and a new cabinet key he had had made. I reimbursed him and he showed me where and how he had stored the data. He had done a good job. He said the supplies had arrived as per schedule. Since he mentioned he'd be dropping by Monday morning I said I'd say goodbye to him then.

August 20, 1963 (Tuesday) - YY telephoned me to report that he was concerned about the supply of paper and developer. It was easier to estimate how much more would be needed to complete the work in Unit Z than in F's office. However, we both agreed on the fact that this next order would be the last one for supplies and if worse came to worse, if there were a few student activities records remaining when the photographic supplies were exhausted, they could be hand-copied. After my conversation with YY I telephoned LL and gave her instructions for typing up the purchase order for 700 sets of 8 1/2 x 11" Dri-Stat paper for $56.00 and 6 quarts of Developer Pre-Mix for $4.50, or a total of $60.50 for delivery to YY on Thursday morning, August 22. I then telephoned Unit Q, and since II was out for the day, gave lengthy instructions to his secretary about calling in the order when II received the requisition in the University mail Wednesday morning.

August 21, 1963 (Wednesday) - YY telephoned me this morning to report that he and OO had completed the contents of 106 folders the previous day and that he would send OO over to F's in the afternoon to begin pulling Student Activity Record Cards. I told him what additional supplies I had ordered for delivery to him on Thursday morning. I then telephoned II to learn if he had received my message from the previous day. He had not received it so I had to repeat the details over again. He promised to take care of everything.

August 22, 1963 (Thursday) - YY telephoned me in the morning to report that things had really been fouled up for him on Wednesday. By pure coincidence T's secretary had placed an emergency supply order with the Irwin-Rheame Company and through a misunderstanding had delivered her order along with ours a day in advance. YY, knowing what I had ordered, refused to accept the other supplies and told them to take them back. Somehow T's secretary learned what he had done and gave him a dressing down I guess he won't soon forget, and in an effort to extricate himself he got involved with II in Unit Q. As a result of the time delay YY was only able to do the contents of 48 folders. We agreed that I would telephone him at home at 10:00 p.m. When I did so he informed me that he had worked hard to complete the remaining 60 folders, the average contents of each one amounting to 20 pages. We made arrangements for me to meet him at 8:30 tomorrow morning with the car to help transport the machine and supplies from Unit Z to F's office. I then telephoned OO and asked her to report in the morning directly to Unit Z to refile the folders YY had finished.

August 23, 1963 (Friday) - I drove down to the University expressly to help YY. I arrived on campus before 8:30 so telephoned the secretary of the business manager of Unit O and arranged for a hand truck to be at the alley loading entrance in twenty minutes. I then met YY and we loaded the things into the car and drove to Unit O. We loaded them onto the hand truck and YY took the truck up the freight elevator to F's office. As soon as I had parked the car I joined him and cleared YY's working location and arrangements with F's secretary. Then I waited while YY hooked up the machine and got it calibrated for making clear copies of the activities records. After introducing him to several other people on the floor I took him downstairs and treated for coffee while we discussed completion of the project. We planned that he should telephone me Tuesday morning, August 27.

August 26, 1963 (Monday) - In reading back through the log, I noted that in the last paragraph of the entry for July 16, I explained that one reason Unit N was scheduled late in our programing was because each student's activities for the previous academic year are manually recorded on his master card during the summer. The clerk who performs this task works slowly. I see I have not made references, however, to the number of times throughout the summer that I have checked with ZZ as to how far along the clerk was in completing the posting. At one time I asked ZZ for her suggestions as to what we could do if the time arrived when we were ready to reproduce the Student Activity Records and the girl wasn't finished. ZZ had suggested that we go ahead and pull the cards and proceed to reproduce them. If the clerk subsequently found that a card was not in the file she was to go to the adjoining office where YY and OO were working and locate the card with their assistance and make the entry. If they had already reproduced the card then they were to write in the appropriate entry on their reproduced copy.

Although YY was to telephone me tomorrow, he called today about noon. He reported that OO had checked about 960 names on the Liberal Arts Fall '59 roster which had resulted in about 420 cards. So far he had reproduced about 200 of these and OO had copied 87. He said that things were not going smoothly because the clerk was pulling cards while OO was trying to do so. He also pointed out that he could reproduce the cards so much faster than she could make entries that if he and OO proceeded to reproduce the cards, they would have a lot of backtracking to do to pick up the current entries. Of course, he was right and I blamed myself for not realizing this would happen. I asked YY if he and OO would object to not working for a few days. He said he'd been wanting to go away over Labor Day anyhow and he was sure it would be agreeable with OO. I asked him to find out when the clerk thought she'd be finished and he said she had said Thursday. "Alright," I said, "tell OO not to return until Friday and to spend her time pulling cards and hand copying that day and September 3 and you report back to work there September 4." We then agreed that since they had pulled only the Fall '59 Liberal Arts roster names that he should give the clerk a copy of that roster to note which persons' cards had further entries and he and OO could write them in on the reproduced copies. Then I asked YY what he and OO could do that afternoon. He said that OO had copying to do which would take the remainder of the afternoon and he said he'd like to go back to our building and double-check the admissions records against the roster and I said that would be fine.

September 3, 1963 (Tuesday) - I returned to campus today and went over to Unit O twice to check with OO that all was going well. She had finished pulling all of the cards.

September 4, 1963 (Wednesday) - YY returned to work and by the end of the day had finished reproducing all of the cards. OO started to refile the original cards back into F's files.

September 5, 1963 (Thursday) - YY met with me this morning to discuss where he should go from here. I should digress here to explain several dilemmas which confronted G and me from the beginning of the data collection period. The original FTIAC rosters which we were given to work from were divided by college and by date of admission. We, therefore, had one for Monteith, Liberal Arts and Monteith Engineers who entered Fall '59, and three more for those students who entered Spring '60 -- a total of six rosters. G and I both agreed that hectograph stencils should be cut and run off in order to supply adequate copies to work and collect from and so it would be possible to package the data collected with a checked-off roster. It was our belief that ultimately the data would be compared by college and admission dates and that it would facilitate the work of those who use the data if they were to be packaged that way. On the other hand G and I supposed that in most offices he would be confronted with straight alphabetical filing, i.e., not divided by colleges and date of admissions. We realized that it would be extremely time-consuming to cut the stencils on a straight alpha listing which would amalgamate the six groups into one, yet to do so one seemingly would expedite the locating of the data. Eventually, however, the data would have to be subdivided again into the six groups. Since there were advantages and disadvantages of either way, I left the choice up to G who chose to have the six rosters divided as they were. Obviously this meant, on occasion, that he had to go through the same file drawers six times instead of once. In some departments it proved even more awkward wherever records were divided on the basis of active and inactive students and in one instance that I recall (Unit V) records were subdivided yet further on the basis of male and female students.

This rather lengthy explanation is indicated at this point because G used to ask me if it would be wise to stop and arrange in roster sequence the records he had photographed in each department and I had told him no. That the vital thing was to get the bulk of the records he could and that someone could later arrange them in proper order. So this seemed to be the next appropriate task for OO and YY to begin.

Another thing which YY asked me was what should he do with the machine which Verifax had loaned to us. My first impression was to have the salesman pick it up at F's office but I thought I'd better first check with T. I telephoned her office and, learning she was in conference, left the message that I wanted to talk to her. It was only a few minutes later when she came upstairs to our office. She wisely advised us that if the salesman appeared in no hurry to have it back, we keep it a while longer in case something unforeseen developed. YY and I also talked briefly with T about eventually doing the student voting registration tabulating. It was agreed that we'll not know what bugs need to be worked out until he has gone down to the City Clerk's office and tried it.

OO spent the rest of this day completing those student activity cards which could be hand-copied with expedience and she also continued the refiling of the original cards.

YY carried the machine and supplies back from F's office by hand. It required him to make several trips but unfortunately I was unable to help him with the car this time. The distance wasn't as great this time, however. After doing that he spent the remainder of the day interfiling the hand-copied student activity cards with those which had been photographed.

September 6, 1963 (Friday) - YY started to make a fresh copy of the roster check lists and to file the cards in proper order. At one point during the morning he came in the office to express his general dissatisfaction with OO because she worked too slowly. He said that he felt quite capable of finishing this phase of the project without her. I urged him to wait a little longer until he was positive of what he was saying. For one thing he felt it was taking her too long to finish over in F's office. She was finished with her refiling over there and came to the office in the afternoon where he was working. In the meantime he had become suspicious in checking through the rosters that she had not pulled the student activity cards for those students whose names (mostly married) occurred on the roster errata which G had provided. Again YY came into my office expressing his dissatisfaction with OO because of what he regarded as her lackadaisical attitude toward the job. He just couldn't understand how, when she knew there was an errata, she wouldn't have remembered to check for the names on it too. At any rate he told her to go back to F's office in the morning to check the married names and to hand-copy any she found.

September 9, 1963 (Monday) - YY was in the office when I arrived this morning. He said he didn't think the data were stored in the cabinet as well as they could be and he wondered if we could buy some cardboard filing boxes which would fit in the cabinet. I suggested he consider cardboard boxes which super-markets discard but I could see from his facial expression that this wasn't good enough. After he left, I got thinking about his request and the more I turned it over in my mind the more I felt that if this is what he thinks he needs to do a top-notch job, then I owe it to him to at least look into the cost. He had said he needed four filing boxes. I telephoned LL to ask if she had any idea of the cost and she said she had four new ones she could give me. I told YY and he was very pleased.

He spent the morning carrying the admissions folders up from downstairs and started to make a check list roster for them and spot-checked the legibility of the high school records reproductions for those students admitted Spring '60.

OO returned from F's office in the afternoon. She had found 22 more student activity cards and I was glad when I learned this because she had not regarded YY's suggestion of checking for married names as potentially fruitful and he thought she'd resented his sending her back there. With the additional cards, he figured out the percentages by the two admission periods and the three groups in the sample.

YY and OO spent the remainder of the day arranging the Admission folders in order.

September 10, 1963 (Tuesday)- YY and OO spent the morning completing the check-list rosters on the admissions folders. In the afternoon they double-checked the transcripts which evidently were already badly mixed up because they had been grouped by letter and fastened with rubber bands. In addition the transcripts of students having "conditional" or "trial" status at the time of admission were grouped separately. They interfiled these transcripts in roster order with the others. During the process they discovered (unbelievable after the checking G had done) that there were four transcripts missing. I telephoned Unit R and they promised to send them out in the mail immediately.

September 11, 1963 (Wednesday)- YY and OO worked hard and steadily again today and I'm impressed with the thoroughness with which YY tackles the job. He is somewhat of a perfectionist in his determination to have everything accurate and this pleases me. They double-checked the 1600 plus transcripts and made a final corrected roster to accompany with the transcripts. The transcripts which I ordered yesterday arrived so that made them complete. Then YY and OO started to plan how they would file the data in the cardboard boxes. They inserted the Admissions folders, transcripts and student activity cards in the file boxes because it was easier to work with them in the boxes as opposed to stacks on the floor. On the 10th YY had discovered there were admissions records missing and he made a list of these with the intention of tracking them down and getting them.

September 12, 1963 (Thursday)- OO went to the Admissions Office this morning and located the missing folders. The contents had to be copied, which took time but they were both pleased that they now were complete. These were then interfiled and appropriate corrections were made on the rosters. During the afternoon they worked further on cutting and labeling dividers for the cardboard file boxes and systematizing the outside label of contents. They did a magnificent job and YY really deserves the credit because OO doesn't have this ingenuity. They then carried the four heavy boxes downstairs to store in the metal cabinet.

September 13, 1963 (Friday)- This morning when YY and I met in my office we agreed that this should be OO, last working day. He had thought he could get along without her earlier in the week but I advised him against our letting her go until he was positive he wouldn't need her. He told me he was glad that she had worked all week, especially because she had been valuable in locating missing data in the offices where she had to return. The two of them spent all day on the Unit V cards. They were all there but distributed among active, inactive, male and female and even found five more student cards of the sample group amongst the extra cards which the Unit V had said we could take because they were throwing them out. So all in all they did a good day's work putting all of these cards in the proper alphabetical order according to roster. These cards were boxed with labelled dividers which would facilitate proper identification. I thanked OO for all of her help and had her fill out her last pay requisition.

September 16, 1963 (Monday)- I don't really know what YY did on the project today because office traffic became exceedingly heavy since registration begins on the 18th. I told YY that I didn't know when we could get together again to plan. He said it was agreeable with him because he needed time to finish up the Science of Society research project he was doing for T's class. We agreed to suspend operations until after registration was over and classes had started.

September 25, 1963 (Wednesday)- During registration I spent some time up in the area where students select their class cards. U walked by me on two different occasions and he was very cool toward me. Since then I have wondered what was bothering him and concluded that it must be that I had never formally thanked him for his cooperation on the project. I dictated a very gushy letter to him with a carbon copy to his supervisor. When I read the typed letter before signing it I had to smile to myself as I remembered a remark a bartender had made one time when I was in Cuba. He was mixing an especially potent drink and when he handed it to the customer he said, "This no do, nothing do!" I thought the same thing about the letter.

YY stopped by the office this afternoon to see if I'd made any arrangements for him to begin on the students who had registered for voting. I was embarrassed to have to tell him I hadn't done so as yet. I'm annoyed with myself about this because I seem to have a block about this and I have no right. I've tried to analyze and have concluded it is either because

I won't know how to answer questions YY will ask of me or else it is because I feel the outcome will not be worthy of the effort. Since voter registration records are kept by each local municipality, I foresee a variety of problems which (maybe because I'm not adept at statistical sampling) would seem to cause inaccuracies on a large scale. Obviously a student has to be 21 years of age to register. We have known that the Monteith student population's median age is lower than Liberal Arts'; many addresses on transcripts are campus (Detroit) when the legal address is likely to be elsewhere but one can't be sure. If we were to use the beginning samples in Fall'59 and Spring '60 even if they are no longer active in the University, then we would not be sure they were still at the original address or still bearing their original name. On the other hand, persons not previously listed as living in Detroit could be now. My mind spins on and on about all of the ramifications of this. It should be made clear that I wholly approve of the idea, but I am skeptical about my being able to assist in the gathering of data which would be accurate, complete and valid. I should stop being so skeptical and give it a try.

September 27, 1963 (Friday)- Had to call U today about scholarships and he was very genial--guess the letter did the trick.

I've been concerned about not having called the Verifax salesman about picking up his photographing machine. I telephoned his office this afternoon, learned he was out, and left the message. Within ten minutes he walked into my office by pure coincidence. I told him we were through and told him how grateful we were for his cooperation, prompt service, and loan of the machine. He said he'd pick up the machine early next week.

I tried to telephone AAA at both her office and home without success. Thought maybe she could give me a lead on who to talk to at the Detroit City Clerk's Office. I then telephoned there directly about 4:15 p.m., but was told they closed at 4:00 p.m.

October 1, 1963 (Tuesday)- Called the City Clerk's Office in the morning and talked to BBB who apparently was CCC's secretary. My conversation with her got rather involved in red tape for possible clearance with the Election Commission. But fortunately she suggested another possibility. While it was true that their office on the 13th floor of the City-County Building had the ward and precinct records with each voter's oath and stamped voting dates and a directory by which by checking an address it would be possible to determine the appropriate precinct, I might call the Director of Elections' Office on the second floor and talk with either his assistants DDD or EEE. I talked with the latter who was very nice. He told me that they had an alpha file of 855,000 names--all the registered voters. The file also gave the ward and precinct for that voter and he said he'd be more than willing to get YY started. He asked about what day YY might be down and I said perhaps by Friday, October 4.

I saw YY shortly afterwards and gave him all the directions, the man's name, etc., and suggested he do a trial run on a group of transcripts which bore Detroit addresses of students who would have attained the age of twenty-one prior to the last City election. He said he'd try to get to it by Friday.

In the afternoon, T telephoned me and I told her about the arrangements I'd made on voter registrations and she seemed pleased. She asked if I'd mind showing her how the data were stored and I said I'd meet her at the chapel. While I was waiting for her on the street, I saw YY and suggested that if he weren't busy perhaps he could help the two of us since he had stored them in the cabinet. T arrived, explaining that it would aid her in discussing the data with a coder if she was familiar with the location of the data. YY proudly explained it to us.

October 7, 1963 (Monday)- Last Friday I didn't see YY and assumed that he must be working downtown with EEE. Around noon today YY's roommate was in my office and I asked him if he saw YY would he tell him I wanted to see him. YY came to see me late in the afternoon and I asked him if he'd been downtown on Friday. He said he was extremely sorry but he had not been able to go. I explained that I should perhaps telephone EEE but would want to indicate a date when he was sure he could be there. YY said that T had told him she wanted him to check each of the 1600-odd admission folders to make certain there was a high school record there and that if it had been photographed that it was legible. He asked me what I wanted to do first. I told him that if T had asked him to do this, then he should. How long would that take him? He said it might take three weeks. I told him this seemed a little long to me and maybe it would go more quickly than he thought and suggested he work at it for a few days and then give me a date when he could go downtown.

In gathering the voting registration data we started with the six (6) FTIAC lists. Because there might be a non-student with the same name, if we went to the Board of Elections with just the name alone we might wind up thinking the student had registered when actually it was someone else with the same name. So we added to the FTIAC lists the student addresses from the transcripts.

Where the student was living away from home, we recorded both his home address and his "campus" address, since he might be registered to vote at either address. In addition, we defined any addresses within a defined area as being a campus address, and checked the student's admissions records to see if the student had listed a different address when admitted. This defined area was determined after a talk with the College Advisor, who was considered most likely to know the residence patterns of the students living on campus. The area defined was bounded on the north by Grand Blvd., on the south by Canfield, on the east by Bush Street, and on the west by Trumbull. If a student had listed a different address when admitted, but now was within this area, we checked both addresses for evidence of registration to vote.

We decided not to attempt to eliminate from our check of voter registration records any students who were not yet 21 years old. Since the group in which we were interested had entered the University in 1959 and since our checking was done in 1964, we felt there would be very few students who would not yet be old enough to register to vote by that time (a student would have had to have entered the University at the age of 16 or younger). Thus, we obtained the student's date of birth from the voter registration records rather than from University documents.

All students with Detroit addresses were checked in the alphabetical file at the Detroit Election Commission. In addition, we sent a student assistant to those communities which were listed as the residence of at least four students which were accessible on public transportation. We sent a form letter to the city or township clerk for all other addresses which were in the area defined by inclusion in the telephone directories of Metropolitan Detroit--Detroit, North Area, West and Downriver Area, and East Area. The form letter was accompanied by a list of the students whose addresses came under that voting jurisdiction, with a request that we be informed if the student had registered to vote, the date of registration, and the date of birth.

The communities where data were gathered by our representative were: Dearborn, Oak Park, Lincoln Park, Royal Oak, East Detroit, Allen Park, Taylor, Grosse Pointe, Grosse Pointe Woods, Grosse Pointe Park, St. Claire Shores, Huntington Woods, Harper Woods, Highland Park, Hamtramck, Warren, Southfield, Hazel Park, Birmingham, Wyandotte, Ferndale, Centerline, Pleasant Ridge, Garden City, Melvindale, Southgate.

The areas from which we gathered data by mail were: River Rouge, Orchard Lake, Farmington Township, Southfield Township, Berkeley, Grosse Pointe Farms, Mount Clemens, Namkin Township, Roseville, Madison Heights, Ecorse, Livonia, Lake Orion, Bloomfield Hills, Romulus Township, Trenton, Grosse Ile., Wayne, Farmington, Bloomfield Township, Flat Rock.

In the original list, this covered all but 43 of the names.

Next we faced up to the problem that many of the women students may have married after our list was prepared and so they would be registered to vote under their married name and perhaps in a different voting jurisdiction.

So, we telephoned the home of all female students for whom we found no record of voter registration and asked if the student was by this time married. If she was, we attempted to obtain her husband's first and last name and her address. If we were told that she was married, but the respondent refused to divulge the married name or address, we coded column 8 as No. 9--"don't know if registered." That is, we had no way of knowing whether she was now registered under her married name in a different voting jurisdiction. If we got her married name and new address, we checked the Detroit addresses in person at the Detroit Election Commission and for all other communities we checked by mail.

After all these data had been gathered, we discovered as a result of work being done on the student activities coding that our original six lists did not include all the names considered to be in the sample under study. Missing were 43 Monteith and 188 L.A. We were provided with two additional lists of names, and we followed the same procedure with these names as we had with the earlier lists except in two respects. First, there were so few names from the suburbs of Detroit that we sent letters to the city and township clerks for all addresses outside of the city of Detroit. This brought responses from the following communities in addition to those listed above: Pontiac, Swartz Creek, Inkster, Utica, Clawson, Allen Park, Greenville, Blissfield, Dearborn Heights, Warren, and Owendale.

Second, by the time we discovered these added names all but a few of the admission records of these students had been destroyed by Unit Z (our personnel had not obtained the admissions records of these particular students). Therefore, we had to rely on the University's Records Office for addresses for these students. This office keeps records only on students who completed at least one quarter of work. Since many of these students did not complete a quarter of work, we were unable to get their addresses and were unable to check to see if they had registered to vote. Those students in their category were almost entirely, if not all, Liberal Arts registrants.

In a number of cases, when we were coding the data we discovered that either the registration date or the birth date was wrong, because the student would have to have registered long before his 21st birthday if both dates were correct. We coded these students as registered, but have held up the further coding of the remaining columns until we have a chance to check.

In Column 7, we classified any female as "not married" unless she was married already while in college or we had information that she was married after leaving. Obviously, this means that the "not married" classification is unreliable to the extent that we were unable to reach by telephone some of the homes of women not registered under their name used in college and that some of them were married later without our knowing it.

In Column 8 we classified as "Don't know if registered" any students who were listed at an address outside of the area which we checked by personal visit or by mail to the board of elections, or who were from areas where the city or township clerk did not return the information we requested.

We should note that it is possible that some students may have registered to vote either before or after entering college and then allowed their registration to lapse. We would not have them recorded as registered to vote, because we were concerned only with those who were registered at the time at which we did the study.

Because it was possible that a father and son (or mother and daughter) might have the same first name and even middle initial, and be living at the same address, we decided that we would do a further check on any name we found registered when the date of birth would indicate someone older than 30 years of age. In these cases, we went back to the admissions records to check the date of birth. If the date of birth matched, we knew that we had the correct information. If the date of birth did not match, we assumed that the person we found registered was a relative of the same name, and that the student was not registered (unless he was one of those students living away from home whose records we were checking in two different voting jurisdictions).

These tedious descriptions describe a tedious and costly process, whose advantages over questionnaire research are now obviously not ones of cost or ease. And while archival records do not have the same weaknesses, artifacts, and irrelevancies, they do have their own imperfections. In no sense are they the "direct truth" or criterion knowledge. In each instance records are kept by persons of heterogeneous motives and perceptions of duty. All we can hope for is that the differences and changes in such record keeping are not confounded with the social treatments or experiments we are trying to evaluate.

This is an interim sample. We are now burdened with a great volume of records, some of which have been coded into performance measures, others of which remain unrefined. In the next few months we should learn whether Monteith made a difference, or at least whether or not these institutional records, stored up for other purposes, have reflected it.

30 Sept. 1965

An Evaluation of Thorndike's Classic Studies of Formal Discipline[1]

Paul C. Rosenblatt

University of Missouri

No one doubts that specific transfer of training occurs as a result of classwork. The doctor is trusted to transfer specific responses acquired in medical school to the clinical situation. The young elementary school teacher is expected to perform specific responses acquired during teacher training. Moreover, no one doubts that various kinds of transfer of training occur in the experimental psychology laboratory.

What is debatable is whether or not training in the classroom produces nonspecific positive transfer. Individuals who believe that nonspecific positive transfer occurs as a result of formal school work claim, in general, that just as the muscles are developed by gymnastic exercise so the mind's faculties are developed by difficult mental work in the classroom. Perhaps without attempting to describe the dynamic mechanisms involved, they might assert that a course like Latin, which requires deep concentration and long periods of fatiguing work, would produce greater prowess in other situations requiring concentration and fatiguing mental effort than would an undemanding course such as cooking.

The most frequently cited of the numerous studies of the effects of formal school discipline are a pair of studies carried out under the direction of Edward L. Thorndike (2, 10). Thousands of high school students were

---

[1] Supported in part by Project C998, Contract 3-20-001, Educational Media Branch, Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act, Donald T. Campbell, Northwestern University, Principal Investigator.

given one form of an intelligence test at the beginning of a school year and another form at the end. Ostensibly, the effects of different courses, holding all other courses constant, on gains of intelligence scores were examined. Broyler, Thorndike, and Woodyward summarize the 1924 report (10) with a statement applicable to their own conclusions: "The general results were that amount of gain bore only a slight relation to the studies taken" (2, p. 377). Is the conclusions warranted by the data?

The Thorndike studies ran into the serious problem of nonrandom distribution of students to "studies." Pretest scores were found to correlate with school programs. The brighter students tended to take academic courses; the less bright students tended to take nonacademic courses. Since brighter students generally show greater gains in intelligence through time even if intervening experience is held constant, more difficult courses could be expected to appear to produce superior gains in intelligence whether or not they actually did. It was, therefore, decided to statistically correct posttest scores by removing the gain which would occur even if no courses were taken. There were four different sources of estimates of natural gain that were used in one or both of the studies. None of these sources can be said to be satisfactory.

The difference in gains between Negroes and whites, who differed initially in intelligence test scores, was used to provide one estimate of natural gain at different levels of intelligence. This method is justifiable only if Negroes and whites could be demonstrated to have similar configurations of intervening course work. But they were not demonstrated to have had similar intervening experiences. It can probably be assumed that they differed in course work in the same ways that students differing in intelligence but not in race differed. Consequently, one source of the estimate of natural gain in intelligence is in error. If one is allowed to speculate, it might well be that around 1920 Negro students were represented in disproportionately large numbers of non-academic courses. If this were true and if academic courses such as Latin and algebra did contribute to gains in intelligence test scores, then it follows that, by using Negro-white differences in gain to correct for natural gain, Thorndike and his co-workers actually removed part of the differences in gain due to course work from their posttest scores.

A similarly condemnable estimate was derived from a median obtained correlation between gain and pretest plus posttest of approximately .09. This was truly an unfortunate choice of evidence, for with such a part [gain] with the whole [pretest plus (pretest plus gain)] correlation, one can obtain a positive correlation even when gain scores are unrelated to pretest scores. There may even have been no reason to correct for natural differences in gain. According to Table XIV of the 1924 study (10, p. 88), the greater the pretest scores the less the gain in intelligence. Possibly formal discipline is less than worthless, or possibly the high school students who had been following an academic program had reached an asymptote in possible improvement as far as the intelligence test used was concerned. Thorndike dismissed the negative correlation between pretest and gain as an error in measurement. However, the negative correlation between pretest and gain suggests that by "correcting" for a positive correlation between pretest and gain Thorndike and his co-workers may have obscured large effects of formal discipline.

Following the same faulty reasoning leading to the erroneous handling of race differences in gains as being purely ability differences, unresponsive to differences in intervening experience, "corrections" were made, in the 1924 study, on the basis of sex differences. However, sex differences were so small in the 1927 study that they were not used in "corrections."

In the 1927 study, one of four estimates of the effect of ability on gain was made by examining gains as a function of scores on a paragraph reading test. Assume that the small sample of subjects used was representative and that the paragraph reading test was correlated positively with the pretest. One could reasonably claim for this new estimate that, to the extent that the better students take academic courses and to the extent that the academic courses effectively discipline "mental faculties," correcting gains on the basis of a correlation of gain with a correlate of pretest scores tends to obscure any real effects of formal discipline on posttest scores.

A second problem faced by Thorndike was that of heterogeneity of programs. No configuration of course work was taken by more than a few students. Thorndike had intended to measure gains attributable to courses by comparing pairs of courses, holding constant all other courses students were taking. At best, very small groups of students with identical programs were available for comparison purposes. Therefore, Thorndike decided to increase comparison group sizes by treating as identical those courses which appeared with similar frequencies at each level of gain, without attempting to remove the effects of other, correlated, courses. In other words, two courses were treated as identical if, for example, both were taken relatively often by students who showed low gains and relatively rarely by students who showed large gains. That two courses might appear to have similar effects as an artifact of the programs in which they were typically embedded was not considered. Moreover, nothing resembling a test of significance of similarity was performed before deciding to combine groups of courses for data analysis purposes.

Utlimately, the "corrected" data yielded evidence that, compared to being bright, taking a difficult program of academic courses is ineffective at producing gains in intelligence. There appeared to be differences in gain between intellectual studies and physical skill studies, but these differences were a fraction of the differences in gain between those individuals of highest initial ability and those of lowest. Because the Thorndike studies did not successfully cope with the problem of differential selection and because the independent variables that were compared were neither independent of one another nor classified independently of their effect on dependent variables, the conclusion of the studies cannot be said to be valid. The hypothesis that formal discipline has effect on intelligence measures was not tested by these two monumental studies. Nonetheless, many recent educational psychology textbooks give the two reports serious consideration. (See, for example, 1, 5, 6, 7, 8, 9).

How might one avoid the differential selection problem and the problems of separating cause from effect and cause from cause? The ideal way is to randomly assign students to treatments thought to be differentially effective. Such a procedure would be feasible in a school system which allows students "electives." The freshmen in a school system might be randomly assigned either to a Latin "elective" or to a world history "elective." One would need to assure that the different treatment groups do not differ in such possibly important things as percentage of opposite sex classmates and time of day of class. If more students prefer one treatment to another, there might be differential profit from the two treatments due to a motivation-treatment interaction. One could suppress the reactive effect of measurement by employing institutionalized intellectual assessments both as checks against biased assignment of students to treatments and as posttest measures (4).

Other advantages of using random assignment exist. One eliminates, by using random assignment, differential mortality (3,4), a problem which Thorndike failed to consider seriously. To the extent that it leads to the elimination of a pretest, random assignment eliminates a pretest-treatment interaction explanation from competition as a valid explanation for positive results. It is to his credit that Thorndike was aware of the possibility of pretest-treatment interaction (10, p. 10) and even spoke of attempting someday to differentiate between the effect of pretest-treatment interaction and the effect of treatment.

Where random assignment of individuals is not feasible, random assignment of classrooms, of schools, or of school systems to treatments may be possible. Where no random assignment is possible, "quasi-experimental designs" (4) may be useful. For example, in certain school settings, a multiple time-series design (3, 4) which would use the test records accumulated for each student over many years, might provide a very strong test of the effectiveness of formal discipline.

It may be that there is an interaction between amount of education and effectiveness of formal discipline such that formal discipline is very effective for students who have no school learning and becomes progressively less effective as amount of school experience increases. To the extent that this hypothesis is tenable, the hypothesis that formal discipline is effective would be more powerfully tested by an investigation at the elementary school level.

Summary

The methods used in E. L. Thorndike's frequently cited studies of the effect of formal school learning on mental test scores are critically examined. Procedures used to compare data from groups of students differing in both initial ability and school program and the procedure used in grouping data are found to be unsatisfactory. It is suggested that studies of the effects of formal school learning might more profitably use either random assignment of treatments or some "quasi-experimental design." It is further suggested that a study of relatively unsophisticated pupils might constitute a more powerful test of the effect of formal discipline than a study of high school students.

References

1. Blair, G. M., Jones, R. S., & Simpson, R. H. Educational Psychology, (2nd ed.) New York: Macmillan, 1962.
2. Broyler, C. R., Thorndike, E. L., & Woodyard, E. A second study of mental discipline in high school studies. J. educ. Psychol., 1927, 18, 377-404.
3. Campbell, D. T. From description to experimentation: interpreting trends as quasi-experiments. In C. I. Harris (Ed.), Problems in Measuring Change. Madison, Wis.: University of Wisconsin Press, 1963.
4. Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963, 171,246.
5. Frandsen, A. N. Educational Psychology, New York: McGraw-Hill, 1961.
6. Mouly, G. J. Psychology for Effective Teaching. New York: Holt, 1960.
7. Pressey, S. L., Robinson, F. P., & Horrocks, J. E. Psychology in Education. New York: Harper, 1959.
8. Skinner, C. E. Educational Psychology. (4th ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1959.
9. Thompson, G. G., Gardner, E. F., & DiVesta, F. J. Educational Psychology. New York: Appleton-Century-Crofts, 1959.
10. Thorndike, E. L. Mental discipline in high school studies. J. educ. Psychol., 1924, 15, 1-22, 83-98.

From B. J. Chandler, Evelyn F. Carlson, Frederick Bertolaet, Carl Byerly, John Lee, Robert Sperber (Eds.), Research Seminar on Teacher Education, Report on Cooperative Research Project No. G-011 supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare. Duplicated, Northwestern University, August, 1963, pp. 75-120. (To be printed by U. S. Office of Education.)

ADMINISTRATIVE EXPERIMENTATION, INSTITUTIONAL RECORDS AND NONREACTIVE MEASURES[1]

Donald T. Campbell
Professor of Psychology
Northwestern University

This talk will first develop the several themes in the title, and then illustrate possibilities and problems through the discussion of several specific designs.

---

[1] This talk is a continuation of and presumes familiarity with the general orientation which Stanley and I presented in our chapter in Gage's *Handbook of Research on Teaching* (Campbell & Stanley, 1963). My work along this line is being supported by the U. S. Office of Education, Media Research Branch, Project C-998, Contract 3-20-001 under provisions of Title VII of the National Defense Education Act. In the oral presentation of the talk, discussion and presentation were deliberately mixed and this format has been preserved here in edited form. Since the names of many discussants were lost, all have been eliminated here except where their particular projects are being discussed.

Administrative Experimentation

Social scientists and educators often end up in the role of passive journalists, recorders of what is going on. This role is quite different from the scientist in the successful sciences, who never tries to describe all aspects of a real situation, but rather experimentally intrudes in a setting where effects can be noted. In undertaking the goal of exhaustive description the social scientist and the educational researcher have taken quite a different role from the role of the successful scientist whose basic approach is to intrude, interrupt, to make a change and to see what happens. We recognize, of course, that experimentation on important issues like how to wean children, how to punish children, etc. is not easily going to be allowed. There are many obstacles to such experimentation. Nor do we often get the opportunity of experimenting with our theories of social structure, organizational effectiveness and the like. Nonetheless, there is excellent reason to believe that unless we, as social scientists, do get into experimenting we are never going to have a science. We are, of course, trying "laboratory" experiments. But when we do such experiments we experiment with those people over whom we have arbitrary power; mainly the undergraduates in our own classes. The kinds of experiments we do are limited by the power we have; namely, ability to hold them captive for one or two hours at the most. Thus our experimental base is of necessity trivial.

A social science that is to deal with important problems must tie-in with people who are doing experimentation at more important levels, who have arbitrary power over more sustained interactions, over larger groups of persons and a wider range of settings. Who is it that has this power? It is you, the administrators. It is the people who are in charge of factories, in charge of training programs, who are in charge of school systems, people whose decisions (whether democratically or arbitrarily arrived at) are put into effect, people who are the gatekeepers (democratically or dictatorially) for abrupt rather than gradual changes. Thus, if we are to have an experimental social science, the social scientist must develop a liaison with the people who have the power. It is not we, but you who have the experimental laboratories, through being at the site, if not in the decision making seat, when abrupt administrative policy changes are made. What we social scientists can do in situations like today is to convince you of the necessity of keeping books on the experiments you make, and organizing your record systems and your publication practices so that you let us know what you tried and how it came out.

The other evening, Dr. Gage and others, including myself, advised you not to postpone taking action until you have done a "survey," not to wait on reforms until the problem had been "researched." When we gave that advice, which I do not want to renege on, we had in mind that type of delegated research in which an outside group of consultants is hired, in which passive descriptive activity goes on, ending up three or four years later with three or four inches of mimeographed reports which nobody has time to read. What I would like to emphasize instead is the need for experimental research in which you go ahead, make changes, use your own best wisdom as to what needs to be done. Do it, but do it in a way in which you learn whether or not you have made a difference. Do it in such a manner that you can both share the results with other practitioners and at the same time improve the sad state of the social sciences by giving us experimental results to feed upon. This is not research as a prerequisite to action but research in measuring the effects of action, research as a bookkeeping on the consequents. This sounds like a number of other slogans we had in the past, such as "action research." I don't want to disown such heritage, although the slogan "field experimentation" may suit better. "Action research" often seemed something advocated by reformers only, and I want research by administrators of all kinds whether they call themselves reformers or not. In actual practice too, the action research movement was usually not accompanied by good enough bookkeeping to ever tell whether or not a difference had been made. But action is the important thing. When we look into the successful sciences we see that action is the principal means of learning. The model of passive descriptive knowing just doesn't hold up when you examine the successful sciences, physics, biology and the like.

Nonreactive Measures and Institutional Records

This has to do with what philosophers call the "scandal of induction," e.g., the fact that you can't prove that the sun will rise tomorrow, and more generally that scientific knowledge cannot be deductively proven. Scientific "knowledge" is instead a logically unjustified extrapolation from previous experience into the future. Hume lived a long time ago but we are still very slow in recognizing his truism in our general thought. The writings of such modern philosophers as Karl Popper (1959) are helping.

In making these brave extrapolations we make certain assumptions about the general nature of the world. The earlier statement of these assumptions, as in John Stuart Mill, was that we assume that nature is orderly. In the more modern version of this assumption, which seems to be more descriptive of our activities as scientists, we don't necessarily assume that the world is uniform but we assume that it is "sticky," i.e., that adjacent parts of the world in time or space are more apt to follow the same laws than are nonadjacent ones. We now recognize that at any given moment the laws of science are imperfect simplifications but their imperfection is greater the farther we extrapolate them. The greater the difference between the area in which a scientific law was developed and the area in which we are trying to apply it, the less confidently we can extrapolate and generalize. The extrapolation is never logically certain but it is more plausible the greater the similarity of conditions, the less distance in time, space and similarity.

This has implications for us when we as educators working in real city schools try to extrapolate the experiments in the college laboratory. If, for example, we have laws of group formation and ego involvement which are based upon one hour of face-to-face interaction among college sophomores meeting a laboratory requirement, and are trying to generalize from these to laws of group formation and ego involvement in a forty-hour week, month after month, we can see that our conditions of similarity are relative low, that the extrapolation is relatively distant and that the plausibility of extrapolating is tenuous. This implies that even though one borrows from the experimental social sciences and their one-hour college sophomore experiments, one has to cross-validate these principles in one's own arena. If in borrowing one needs to cross-validate with administrative experiments anyway, why don't we just start experimenting where we live? If we are doing experiments with samples of an administrative principle, we are thus providing a basis of experimentation that is close to home, where the extrapolation is small, and where the likelihood of valid extrapolation is great.

One of the features which makes an applied living situation much different from a laboratory experiment is something generally recognized under the name of "artificiality." Such artificiality exists when one gives college sophomores an hour of soul searching questionnaires introducing lots of issues that they have never thought of before and gives them an hour of experimental interaction, and then another one of these questionnaires. In doing research in the school system, typical Ph.D. candidates may come in with enough measuring instruments to bore any student or teacher, possibly with unusual types of questions, and with a clear interruption of the routine. Such measurement procedures create a situation which is very different from the situation which we are trying to generalize to. For an example away from our immediate questions, suppose we are trying to estimate the effect of a movie such as "Gentlemen's Agreement" (an antiprejudice movie) for people who see it under normal theater conditions. Suppose we precede this movie with a 100-

item soul-searcher on antisemitism. An audience with that kind of priming is quite different from the audience we are trying to generalize to. It has been so "warmed up" that it could not possibly see that movie, as some women have, without recognizing its social problem characteristics. But some natural audiences have been so preoccupied with the love story of this beautiful girl and her quixotic boyfriend that they didn't notice that one of the dimensions of the boy's quixotic behavior included registering at hotels under Jewish names, etc. Thus, one of the dimensions in which a so-called experimental research may differ from the settings we would like to generalize to is through the use of "reactive measurement devices" (e.g., Campbell & Stanley, 1963, pp. 179, 188). Measurement devices which are strange, alerting or make people feel like they are being guinea pigs, instruments which start thought processes and problem solving activities, are all apt to be reactive, i.e., instruments which cause change as well as measuring change.

Many kinds of measuring devices may be regarded as reactive in this sense. We can take a journalist as a measuring device. If we introduced into this very conference a man who has a big badge saying "Press" on it, it would probably alter this discussion. Such a newspaper reporter is a reactive measurement device in that he changes what he intended to measure. A simple thing like scales that measure a woman's weight can be a reactive device. Consider an experiment on weight in which everybody has to weigh in at the doctor's office every morning. This in itself would cause some to reduce no matter what kind of pills were being given them; the scales would be a source of change.

Thus one of the things we should try to do in experimentation for applied purposes is to use measures which are as little intrusive as possible. Measures could be reactive in some situations which are not reactive in others. If I were to use a vocabulary test in an experiment in the effects of television for the general population, this vocabulary test for the average adult would be reactive, in that it might initiate new efforts at learning vocabulary. It would be better in this setting to, unbeknownst, take a thousand word sample of a conversation to be later scored on vocabularly level. But a vocabulary test in the school system is not a reactive measure because these students have become so used to being continually given tests; testing is a part of the regular bureaucracy. Thus, one of the ways of being alert to the possibility of nonreactive measures is to look at those things that we are already keeping book on, those institutional records which are already a part of the natural situation into which we wish to generalize. Such is the second theme of the title, an emphasis on using measures of effect which are archival or are part of the institution we want to generalize to.

An Ideology of Continuous, Contagious Institutional Experimentation

As a final introductory point before we get into specific design problems, I would like to discuss the strategy for getting research done and an ideology of institutional experimentation. If we can be alert to the experimental opportunities that are occurring every day, if we can anticipate those ways of putting administrative decisions into effect that will enable us to evaluate our results, we can establish experimentation as something that is a regular part of our institutional bookkeeping, something to be done continually and repeatedly, rather than waiting around for the big one-shot research grant. The use of institutional records will cut our research costs greatly. This kind of research reality testing can be made a part of our regular budget, so that we don't wait for the big projects as for an outside deus ex machina that will come and save everything. In passing it can be noted that big experiments very often have small outcomes. There is perhaps something wrong or wasteful in the strategy of big experiments. If we look again at the successful sciences, we find that their truths are buttressed by hundreds or thousands of replications. Every important experiment becomes built into the measuring instruments of the next level of science, so that the major phenomena are replicated again and again in every laboratory. In the social sciences and in education, we have the notion that one experiment will do it and we invest large sums in single big experiments, which after all only test one thing, in one setting, once.

What we need instead of few big experiments is a contagion of imitative experimentation in which everybody is trying something continually. If anybody hits on anything hot, then this is cross-validated by being imitated by many other people who continue to keep records on it to see whether it really does work. Now, in my field, many hot results turn out to be false positives. (We use the five per cent level of significance and thus of every 20 experiments we do at least one looks interesting, but this would happen even if all we were dealing with were pure chance.) The experimental procedure, the bookkeeping, the reality testing, needs to be continued by those who borrow a hot idea as well as those who innovate it. We need to continue this bookkeeping and it is in the cumulative and public record of such experiments that a viable science or applied science can grow.

In a similar way, I don't think that school systems should wait around for a Ph.D. candidate to work on its problems. Ph.D. candidates are even less dependable than large grants, because they usually have to get their job done in a year. As a matter of fact, a six months span of observation is long, and dissertations tend to be one-shot affairs with the man who did the research hired some [illegible]ve-thousand miles away by the time any real effects might be expected.

This ideology of experimentation can be effective within large systems only if we allow heterogeneity within the system, so that each district, or each school, or each classroom can be trying something different, generating something like a free enterprise system in explorations of novel approaches. With the kind of bookkeeping I am going to talk about, with continuous reality testing followed by contagious imitation provides the cross-validation. It is an essential part of the physical sciences in which bandwagon effects of problem popularity either confirm or erase every promising lead. The bandwagon is a part of the scientific process because it provides the strong cross-validation. It is not to be confused with the bandwagon fads in education in which everyone starts applying something with the blind faith that it has been proven elsewhere. Here you are no longer doing reality testing, you have no longer any chance to learn whether the practice is indeed effective.

The cost of such an experimental ideology is the administrative complexity which comes from nonuniformity (which can seem to be a lack of discipline), and the feeling that if students are being treated differently, some of them are getting a less effective mode of teaching than others. These are primarily seen as costs because one has an exaggerated faith in his best policy. A little experimentation would usually show that this faith is exaggerated, that there is no mode of education being used that is so superior that it makes much difference that some students have a variant which someone guesses to be better. In addition, the higher teacher morale which comes from disrupting habituated teachers, and the implicit or explicit competition for excellence which reality-testing brings are gains over-and-above the trial-and-error discovery of better methods.

Intermediate Discussion

We have concluded the presentation of the general strategy advocated. The remainder of the talk will be devoted to concrete illustrations and the experimental design problems and opportunities which they present. But before we go into that, let us break for discussion of the issues raised so far.

Question: You suggest that we use our present records as data for what is going on. This sort of gives us a serendipity of records which we haven't even thought about or looked at for this purpose. I am rather intrigued with this idea.

Comment: This is the major emphasis, although we may want to suggest a few more records to be made institutional just for this purpose. Thus we might want to suggest an annual morale ballot for every student, and an annual progress report from every teacher.

Question: You wouldn't suggest expanding our records to keep data which we don't see any use for?

Comment: No, but if we have the orientation that we are continually going to be evaluating every administrative change we put in, every time we shift books, every time we introduce a teaching machine, every time we send a teacher to a special summer school in-service training, in the end we will seek out the kind of records that are optimal as long as the costs don't get overwhelming. What I am arguing against is introducing different new and unrepresentative measures for each specific experiment. Before we do this let's examine the relevant records that are already available case by case.

Question: You aren't talking about "anecdotal" records. You are talking about more specific and regular measures.

Comment: The type of records I had in mind were things like grades, examination scores, absences, dropout rates, disciplinary infractions, nurses notes, whatever you have, or could have without having to go to the Ford Foundation for a million dollars.

Question: Extrapolation in the area of social change depends on similarity. Then the question can be asked about relative differences or similarities within the bookkeeping system. This is one problem that I see. While you are right in that there are vast similarities between the school system of Chicago and New York, etc., within certain limits, for example, issues related to segregation and economic deprivation, they may in fact be quite distant and this may cause some problem in extrapolation.

Comment: God hasn't given us a promise in advance that science is possible. We have to discover the extent to which it is. I would like to make two responses to the question. First, it is easier to extrapolate from a 40-hour week experiment in a school system in Chicago to a corresponding neighborhood in New York, than it is from my small group research with Northwestern sophomores talking to each other for an hour. Second, with contagious multiple experimentation, what we are going to do is experiment in extrapolation, we will explore our extrapolation ability. For example, Professor Gage suggested that the kinds of teachers that have high morale classes might differ as a function of social class. We certainly will often find "interaction effects" or qualifier relationships, but we will find these by your experimenting in your school system. It will probably turn out that we can extrapolate more confidently where the situation is most comparable. Under the administrative research strategy we would have a pool of research findings in which one could search out the most comparable findings. If the ideology really took hold we would need a journal which would publish research and replication research no matter how the results came out. This would provide us with a realistic sample of our collective experience rather than the misleading rosy-glow of selective presentation of occasional happy outcomes.

Question: One more comment, if you are interested in the question of being able to extrapolate, you are probably also suggesting that in the process of looking at extrapolations ways might be discovered of lawfully taking into account certain factors which prevent simple extrapolation.

Comment: We hope that there are laws to be learned. We hope that there are dependable generalizations to be learned. We cannot tell without trying and we cannot wait until such generalizations are developed in the laboratories of experimental sociology, we have to develop them in our own field.

Question: When one replicates an experiment and it doesn't work a second time, one begins to think that the basic hypothesis is wrong. May it not be that a more fruitful avenue of research is to see wherein the situations were different in systematic ways?

Comment: This is the correct strategy providing it does not degenerate into one-shot apologies for an inadequate law. One should take the hypothesized source of dissimilarity and get someone else to try it out, so that you cross-validate the guess. Both in the laboratory and in field research we have the misleading notion of once-and-for-all experimentation. What we need is the model of cumulative, iterative, trial-and-error processes in which we make each application a real trial and error because in each situation we really evaluate. The publication bias is tremendous on reports of experiments. Only the "encouraging" outcomes tend to be published. Most educational experiments are reported at the crest of the wave before the results actually come in, and the ones that get reported are the ones that sound exciting. We need instead record-keeping such that we take a given "cure" and can get a realistic notion of how it has worked in practice in terms of hundreds of trials.

Question: One problem that bothers me is that most of the records deal with counting strikingly tangible objects, and usually these are the things that are the least important.

Comment: This may be a reason for putting in new measures, such as an annual student morale ballot in which we ask each student to report on where he thinks he's going in life, what kind of a job he thinks he is going to have, whether or not he is going to college, whether he thinks life is getting better for the average man. We need records like this, but records so regular that they aren't seen as introduced just because of a conspicuous new experiment. The issue of validity runs right across the board, and is just as much of a problem in the kinds of measures used in laboratory experiments. Questionnaires, interviews, etc., have the problem of validity just as do attendance records. There are no intrinsically valid measures for anything. We only know indirectly. This indirectness can be found in the kinds of coincidences that are exploited in the visual system (Campbell, 1959). Analysis of any physicist's instrument, such as the cloud chamber, shows the measure to be indirect. The cloud chamber shows some residual products of the beta partical that went past--it doesn't show the beta partical. All measurement is by symptoms. We want to get as many triangulating symptom sources as possible (e.g., Campbell & Fiske, 1959; Campbell, 1960) but for today I would like to emphasize the symptoms that are already available but are not being used. Many times these symptoms are usable even if they have a constant bias and many times it is better to live with a biased symptom than to change the bookkeeping. For example, we have a local problem in Chicago--has Orlando Wilson had an effect on crime rates? Wilson has covered his trail so effectively that we'll never know, because when he came and instituted his reforms he changed the bookkeeping simultaneously, so we have no records that are comparable. Many times it is better to live with a stably biased measurement procedure than to shift to a better one if the shift would be simultaneous with an administrative change you would like to evaluate.

In some sense we will only learn about the validity of these measures through using them in experimentation. Let us take two situations for using the morale ballot. One, we pick 20 students out of the school and say "You are going to be our pilot project." We set them up in a different classroom with a new teacher brought in for the purpose and we give them a morale ballot every week. Compare that with a situation in which every student has filled out a one-page morale ballot every year for a number of years. They are used to it and nobody in the room feels that he is a part of a pilot project; they don't even know that experimenting is going on. Then we might start to measure the effects of in-service group dynamics training given to certain teachers.

I would not make a special case for uniform bookkeeping between school systems. I think I would argue that comparisons between Chicago and San Francisco differ on so many dimensions that it is not an interpretable comparison. I am talking about experimentation within a school system in which you use your records to compare an experimental and control group, or now and then, within one's own system. If you do have a mode of teaching that makes students attend better, it will show up on several ways of scoring attendance. The emphasis is on comparisons with a system but this cannot be made unless you have some kind of bookkeeping that you keep comparable from year to year.

Question: Haven't you been unduly hard on big projects? Isn't one of the problems in education the small amount of research money that has been invested. Don't too many educators draw too much confidence from isolated small studies?

Comment: I am not against big projects. We all favor action and big projects often make action possible. What I am against is this notion that we cannot experiment without big projects. I think each teacher in his classroom can experiment (if he reads the chapter by Campbell and Stanley, 1963, where statistics and design are carefully segregated). For example, he can experiment in changing his teaching method for parallel lists of vocabulary words. There is no need to wait for a big project to do experimentation and perhaps we ought to save the big projects for principles that look really hot as a result of the trial and error of small experimentation. We will come upon some things that seem eminently worth doing, and that can't be done without outside support. But let's not say that experimentation is for the future day when we have big money, or that it is only for that other person who, because he is teaching teachers, must be an expert in research. You in the field know more about the problems, more about the probably efficacious treatments, and should be the persons better able to do the reality testing.

ANALYSIS OF SPECIFIC DESIGNS

In the following sections we will discuss several specific designs involving action programs representative of the problems of this conference. These will include the Rivlin plan for team teaching, Haubrich's problem of the effect of poverty, the Buffalo experiment in revitalizing a low-morale teaching staff, Mrs. Carlson's question as to the difficult-easy vs. easy-difficult order for intern experience and the Northwestern Tutoring Project. The discussion can proceed more effectively if I first put before you the entire check-off list of threats to the validity of experiments (Campbell, 1957; Campbell & Stanley, 1963), some of which, such as reactive measures and reactive arrangements, we have already touched upon.

Check-List of Threats to Internal and External Validity

Fundamental to this listing is a distinction between *internal validity* and *external validity*. *Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific experimental in--stance? *External validity* asks the question of *generalizability*: To what populations, settings, treatment variables, and measurement variables can this effect be generalized? Both types of criteria are obviously important, even though they are frequently at odds, in that features increasing one may jeopardize the other. While *internal validity* is the *sine qua non*, the while the question of *external validity*, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal.

Relevant to *internal validity*, there are eight different classes of extraneous variables which, if not controlled in the experimental design, might produce effects confounded with the effect of the experimental stimulus.

These are:

1. History: the other specific events occurring between the first and second measurement in addition to the experimental variable.

2. Maturation: processes within the respondents operating as a function of the passage of time per se (not specific to the particular events), including growing older, growing hungrier, growing tireder and the like.

3. Testing: the effects of taking a test upon the scores of a second testing.

4. Instrumentation: in which changes in the calibration of a measuring instrument or changes in the observers or scorers used may produce changes in the obtained measurements.

5. Statistical regression: operating where groups have been selected on the basis of their extreme scores.

6. Selection: biases resulting in differential recruitment of respondents for the comparison groups.

7. Experimental mortality: the differential loss of respondents from the comparison groups.

8. Selection-maturation interaction, etc.: in certain of the multiple-group quasi-experimental designs, such as the nonequivalent control group design, such interaction is confounded with, i.e., might be mistaken for, the effect of the experimental variable.

The factors jeopardizing external validity or representativeness are:

9. The reactive or interaction effect of testing, in which, as we've already discussed, a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable and thus make the results obtained for a pretested population unrepresentative of the effects of the experimental variable for the unpretested universe from which the experimental respondents were selected.

10. Interaction effects between selection bias and the experimental variable.

11. Reactive effects of experimental arrangements, which would preclude generalization about the effect of the experimental variable for persons being exposed to it in nonexperimental settings.

12. Multiple-treatment interference, a problem wherever multiple treatments are applied to the same respondents, and a particular problem for one-group designs involving equivalent time-samples or equivalent materials samples.

The Rivlin Team-Teaching Proposal

As the first of our concrete experiments for discussion, let's take one described by Dean Harry H. Rivlin in his paper "Teachers for the Schools in our Big Cities" (Mimeo, October 1962) as distributed to the members of this conference. On pages 14-16 he proposes a plan for using team teaching and related devices to prepare and recruit prospective teachers for challenging urban areas. Briefly, the experimental treatment would be to introduce in a teachers college in a metropolitan area a program of school and community service in lieu of some on-campus courses. The program would start at the sophomore level. Students would be paid as for a part-time job. In the senior year the commitment would be three hours per day as an assistant teacher. Subsequently they would be full time interns.

Let us suppose that we are interested in the effect of this program on the percentages of graduating teachers choosing to teach in the difficult urban areas. Let us first consider the most usual, but least interpretable, condition under which the program would be introduced as an across-the-board change (rather than staged so that contemporaneous experimental and control group comparisons were possible). If the first class to graduate under these conditions is 1966, we might find that 9 per cent choose the problem schools, whereas in the class of 1965, only 2 per cent made such choices. Such an outcome would be very encouraging, yet as an experimental design it leaves many threats to validity uncontrolled. For example, it is possible that the exceptional figure is not the 1966 9 per cent but rather the 1965 2 per cent. (Professor Lawrence Ross and I are analyzing the effect of Governor Ribicoff's crackdown on speeding in Connecticut. It turns out that the most exceptional year for auto deaths is not the good year following the crackdown, but rather the exceptionally bad year preceding.) We gain a great deal from having an extended time series, for example, if the years 1960-66 were to have run in the fashion 2, 3, 2, 2, 1, 2, 9, we would have little reluctance to infer a genuine shift, while if the series had run 8, 3, 5, 9, 4, 2, 9, the same jump from 2 before to 9 after the change would have seemed meaningless.

Consider the measure of volunteering. If we had never asked people to volunteer for such schools before the experimental year, then we couldn't do time series. This should be kept in mind in planning. When we start this Rivlin experiment, it is going to take about a year's preparation before the first group of sophomores begin. This plus the years before the first class graduates gives us some four possible years of pre-experimental data for the time series. In the initial planning year, we want to look at what kind of records are available. We might just for ease of bookkeeping, look at our forms whereby we solicit teachers from the graduating class and see if these can be modified to be more useful for our purposes.

Assuming we get an impressive successful time series such as the 2, 3, 2, 2, 1, 2, 9, how valid is this as an experiment? There are a number of the threats to validity left uncontrolled. But the strategy recommended is not to throw up one's hands in rejection of the invalid, uninterpretable data, but rather to examine the uncontrolled variables one by one to see how plausible they are in this specific instance. Where a plausible rival hypothesis exists, one then tries to make other data analyses that might rule out this rival hypothesis. Looking at the check list, we note that "history" is uncontrolled. That is, some other simultaneous change, not a part of our program, might have caused the effect. How plausible is this in this case? If we learn that the school board, beginning in the year 1965-66 advertised a $1,000 per year hardship bonus for such schools this might seem a plausible rival. We might learn of other plausible rival hypotheses by asking the volunteers why they made the choice they did. (This might be incorporated in the basic recruitment form used in all of the years.)

Maturation and testing do not seem plausible rival hypotheses, in that each year's data are from a separate set of student teachers. An instrumentation bias might occur if our mode of recording changed: if for example, older data were based on those actually placed in such jobs, and our 1966 data were based on all those who asked to be considered for the jobs. Selection is an uncontrolled threat to validity if it turns out that independently of this experiment, the teacher's college started getting a very different type of entrant in the entering class graduating in 1966. High school records for the series of classes could be examined to check this. Not all differential selection and differential drop-out effects would be invalidating, however. It might turn out that the team-teaching and service apprenticeship attracted a group of students who previously had not gone into, or not stayed in, teacher training. Such an effect would be valuable, even though one would want to do analyses to separate it from the effect of changing attitudes in a relatively unchanging talent pool.

The design we have been considering is the "Interrupted Time Series" (Campbell & Stanley, 1963). One would probably want to strengthen it by looking at a similar time series of recruitment data from some other teacher-training institution sending student teachers into the same schools, but not introducing this innovation this particular year. This would produce the Multiple Time-Series Design. Even if the other institution had a quite different base rate for supplying teachers, if it failed to show the 1965-66 jump this would help rule out explanation in terms of general shifts in the teacher employment market, etc. In a still more elaborate experiment, one might divide that first sophomore class into an experimental group, offered these opportunities, and a control group excluded from them. If the apprenticeship billets were in genuine short supply, one might consider allocating them by lottery, optimal for both democratic and design purposes. However, differential treatment often creates great administrative costs as well as "reactive arrangements" in which experimental or control groups may resent their deprivation or arduous duties. Such considerations might make one settle for the quasi-experimental time series. But, if the positions are in short supply, it would be criminal to fill them on the sole basis of eagerness or special privilege: better than twice the number of volunteers be recruited as there are billets, and these then be assigned by lot into experimental and control groups.

We would, of course, want time-series data on many other features than per cent recruitment. Once on the job, we would want data on requests for transfer, quitting the school system, and probably on an annual report form in which each teacher would report on areas of success and disappointment, his long-range goals, his estimates of the chances of the underprivileged bettering themselves, etc., collected, of course, for the pre-experimental class years too.

Question: Why have't you said anything about the necessity for having persons trained in research to give the specific and replicable definition of the treatment. For example, when Dr. Rivlin sets this atmosphere up for his candidates, I can see that there might be variations in how they are received in the schools, and in where they might go. And all of this could produce a long list of crucial variations, anyone of which might cause the opposite results or lead to negative results when tried in other places. What I'm really asking is, why you have not built into the ideology of administrative research the necessity for competence in research? Why have you not called for more rigor and safeguard? Without this I have serious question about the utility of greatly increasing so-called experimentation.

Comment: Who is to decide on research methodology? Look at the list of twelve sources of invalidity of experiments. These did not come from any philosopher of science, these came instead from the criticism of specific experiments by people who were in a position to know of specific weaknesses. My list of factors are abstractions from a literature in which specific experiments have been criticized. In our present case, who is in a position to provide this criticism? I think it is the teachers and the school administrators. In every science we have learned about experimentation by doing it. It is not enough to get more people in to research design courses. The basic cure is to get people doing experiments and criticizing their experimentation in terms of the specific issues that the setting produces. Better than incompetent people try to carry out and interpret experiments than that experimentation be postponed until a year of training in research is given to every teacher. I do seriously believe this. It will help to read the Campbell and Stanley chapter (1963) only because that chapter is an accumulation of types of criticisms, and gives a list of things that have previously gone wrong.

This is not to discourage extension of research training. I think it would be good to give each teacher a course on experimenting with teaching methods in his own classroom practice, and each educational administrator a course on interpreting administrative changes as experiments.

Question: Most of us come from school systems with some repeated testing program. For example, we have accumulative cards for each student, and all testing scores are recorded on it, right through from the elementary grades to the teachers college. At any given point we can go back to look at these records. However, when we go back, we often discover that our data are meaningless because we have not all been using the same tests. One school used Stanford, another California, etc., at the same time, and in addition, each have shifted from year to year. Now within the last four years our own school system has moved to a common test because we can get interesting results. Of course it is going to be a while before we have the same tests for all different schools and systems. My question is, with all of the different tests, how can we have useful time-series data?

Comment: I certainly would be a strong advocate of using as comparable measures as possible within a system. For example, even if you feel that a given test is out of date, it may be better to keep with that inadequate test so that you have for time-series experiments a succession of the average performance of the fourth graders year after year, for example. In defending the utility of such consistent inadequacy, I would like to emphasize the "plausible rival hypotheses" strategy. If you take a modern course in research design you end up with the impression that all experiments are invalid until they are proven valid. Now, the strategy I am advocating is the opposite one: All experiments are valid until you find a plausible rival hypothesis that explains away the results. Thus even though you have not used randomization, even though this and that went wrong, I am asking you to interpret the experiment as valid unless these uncontrolled factors provide a plausible explanation for the way the results came out. Now if a shift in the test administered coincides with a jump in the time series, this shift could explain the result, for we know that test constructors never use the same or really equivalent standardization populations. If, on the other hand, we had test heterogeneity within our school system, this would fail to explain away a shift in mean achievement scores unless there had been simultaneously a shift in the proportion of the students getting each test. Further, it would often be possible to partial out separate time-series evidence for the separate tests. (Another very subtle problem in time-series experiments following the same group of children over a number of years is that the standardization populations for different age levels are rarely precisely comparable, producing artifactual jumps in the time series.)

Haubrich's Question as to the Effects of Poverty on School Attendance

An issue raised by Dr. Vernon Haubrich in one of our discussions offers an opportunity to present some important issues in administrative experimentation. Haubrich's question and observations were on the effects of poverty on school attendance.

One approach to this problem would be the correlational one, comparing absenteeism for various levels of parental income. While such studies are informative, they lack the delineation of directional relationship which experiments and quasi-experiments can offer. The familiar customs of dilatoriness could have caused the poverty, for example. But how can we experiment with poverty?

Haubrich's initial data are in the form of comparisons between absences on days when the school is collecting money from the pupils and days when no such collection takes place. (This approximates the Equivalent Time-Samples Design of Campbell and Stanley, 1963.) If the occasions of collection are not correlated with other occasions inclining to high absenteeism (such as perhaps, Mondays or Fridays), such data can provide a very impressive quasi-experimental design.

One way of analyzing such data would be in terms of the correlation of two time series, after the manner of the economists. One time series is absences, the other is the amount of money which the child is asked to bring (ranging from a mode at zero to some higher figure). This form of analysis, with the exploration of the effect of lagging one variable upon the other, is a mode of analysis which the other social sciences could well borrow from economics (viz. Campbell, in press).

Stating the analysis in this form calls attention to the fact that technically the evidence is correlational, rather than experimental, as long as the occurrence of the "experimental variable," the money collection, has not been scheduled by the experimenter so as to make the occasions totally uncorrelated with any lawful antecedents, including anticipated attendance. Thus one might want to verify the effect by scheduling collection days through a table of random numbers.

Here again, however, the nonexperimental data should not be discarded unless there can be found a plausible rival hypothesis to which one is willing to give the status of a law of strength comparable to one's favored causal hypothesis. For example, were the correlation found to be the-larger-the-collection-the-higher-the-attendance, then one might hypothesize that collections were being scheduled to occur on days of predictably high attendance on the basis of weekly cycles or other factors. In such a case, cause would be from absenteeism (anticipated) to collection. But for the correlation the-larger-the-collection-the-lower-the-attendance, such a reverse causation based upon anticipation is probably not plausible, and therefore the data would be unequivocally interpreted as confirming an effect of collections upon absenteeism. The agricultural economists in correlating annual rainfall with wheat prices have had a similarly unambiguous causal interpretation up until recently. However, were artificial seeding of rainclouds to become more practicable, then a causal direction from (anticipated) wheat prices to rainfall might be plausible.

On a problem such as this, one could look for other "natural experiments," instances into which in the life continuity of a child poverty has intruded at a specific point, instances of catastrophic poverty such as cases where the child is poor because the wage earner has died, or in which an entire factory closed down and moved out of the vicinity, so that poverty is introduced abruptly into a time series. Or one can move from natural experiments into administrative experimentation. Haubrich suggests such an experiment in which the school stops asking the pupils for money. Such an experiment is elaborated in one of the research projects proposed by Group III presented elsewhere in this report.

The Buffalo Experiment in School Revitalization: The Strategy of Staged Introduction

In our discussions here, Doctors Frank J. Dressler and Warren Button have described a program under which the teaching staff morale of a slum school was greatly improved through a variety of means. Let us interpret the Buffalo experiment as a Multiple Time-Series Experiment. What institutional records would be usable? You who have detailed familiarity with the setting are in the best position to be creative as to what institutional records might be tried. Let us suppose that we have measures of average length of tenure of teachers in each school, or the average number of years that they have been on the present staff, or some other indication of turnover rate. We might also have the number of requests of transfer per teacher, and other useful indications are probably in the records right now. Then at a given point in time, we introduce the Buffalo treatment for one of these schools. From the reported impressions as to what happened, we can imagine that over the years covered in the time series, things are getting worse, so that year after year the average length of tenure is going down and the number of requests for transfer is going up. We plot such trends for each school, and note that for this one experimental school we get a reversal of trend, while for the others, the "control" schools, the worsening trend continues. It would be very instructive to have actual time-series plots on each of a number of already available indices. I doubt if this has been done, but it could be, and should be before we offer the experience as inspiration to other school systems.

To use this as a take-off for more general design issues, it can be noted that there is one big disadvantage to always picking the worst group as your experimental group. The Chicago Youth Development Experiment here in Chicago is finding this. We have a social system which runs around putting out fires. The school you will pick as worst is probably going to be in a neighborhood in which many other civic agencies are putting in their strongest efforts. By the time you have been alerted to the fact that this was the worst school, other firemen are already there trying to help things. When we then look at the other events that occurred here in this neighborhood that could have produced this effect, we are apt to have many plausible rival hypotheses, although in this case I don't think it is likely that they would have had direct or immediate effect on teacher tenure.

The Buffalo experiment involved a complex package. On this, I would like to enter a slight disagreement with what Nathan Gage said about doing science with pure variables vs. complex treatments. I think the natural history of experimentation is always to use complex variables first, and, if you get a dramatic effect, then to start further experimentation to find out what part is essential. There is no objection to using a complicated package like your Buffalo package. The very nature of remedial action in social situations necessitate this. If it has a dramatic effect then it is time enough to start peeling it away to find what is the most essential part of it. No matter how "pure" your treatment variable is, it will eventually in the history of science turn out to be a package anyway. And it often turns out that our early guesses were wrong as to what part of the package was causing the effect. If this is so in pure science, it is particularly so if we are interested in applied experimentation. Let's be willing to experiment with packages even though if we get dramatic effects and disagreements as to what part caused the effect, we will have to follow up with more analytic experimental treatments.

The Buffalo study illustrates the desirable strategy of staged innovation. Where an administrator can do so, I would favor staged introduction of innovations, so that one can get something like experimental and control groups for analysis. If we are introducing a new policy, it probably is going to be so cumbersome and to require such supervision from the central office that it cannot be introduced simultaneously anyway. If so, let's introduce it in a deliberately staged way so that we can learn from our experience. If we have a new policy we are introducing, let's introduce it in such a way that the groups that get it first and the groups that get it last are as comparable as possible. All groups eventually get it, but within this transition period the experimental group and the control group can be comparable. The costs of such designs are not great, although it often means a little bit more bookkeeping in the beginning. The schools that are going to get it first should be scattered over all of the city, as should the schools that are second and third, etc., which causes more administrative effort than introducing it in region by region. In the latter case, our first group might be in the suburbs, the second group in the slums, and thus lack optimal comparability. Even if we are going to introduce something as trivial as new texts, if we want to evaluate it let's introduce it by states (perhaps by developing matched pairs or triads of schools and then randomly assigning within these). In the Campbell and Stanley (1963) chapter a distinction is made between quasi-experimental designs and true experiments (such as their designs 4, 5 and 6). It should be emphasized here that administrative experiments can be true experiments and should be whenever possible. Quasi-experimental analyses should be used where nothing better is possible, but our emphasis upon this *should not detract attention to the frequent*, low cost possibility of true experimentation which advance planning can make possible.

The Northwestern Tutoring Project

Over the past few years, a volunteer group of Northwestern University undergraduates, with the advice of Professor B. J. Chandler among others, have been tutoring academically retarded students from several central city schools. Currently two students, Frances Willis and Lionel Dreade, are planning honors theses evaluating the effect of the tutoring upon the pupil tutees. A discussion of the problems of research design encountered offers opportunity to illustrate a number of important points.

First, as to measuring instruments: the initial thought was to introduce personality tests and morale inventories. But as these would almost certainly have to be administered by the tutors themselves, changes might well be due entirely to the increased familiarity with and friendship for the tutors. No such changes could be expected from a control group, nor were they likely even to submit to the tedious testing involved. Differential refusal rates might well explain the differences. It became obvious that meaningful measures would have to be administered by a neutral third party in equally good contact with both experimental and control group. The third party available is obviously the school teacher, with examination grades, course grades, standardized achievement tests, absenteeism records, tardiness records, etc. These institutional records are obviously relevant to the goals of the tutoring project. Although the tutors feel that some of their major impact is upon such intangibles as hopes for self-betterment, trust in the goodness of the majority community, and the like, these will probably have to go untested until the annual or semiannual student morale ballot can be introduced into the classrooms from which the students come. In the meantime, the available records are more than adequate to justify the experimental analysis.

If school records are to be used, why not measure the effects of the tutoring already done in previous years? This would be worthwhile. It would be good to have comparative data on tutored and untutored students from the same classrooms across the several marking periods of the year of tutoring, the previous year, and the subsequent year. However, there is one very plausible rival hypothesis left unchecked under this design, number eight in our list of sources of internal invalidity, the "interaction of selection and maturation." In the previous years, the pupils have volunteered for tutoring, and have had to go to considerable effort to get to the place of tutoring, away from the school building. Thus those exposed to the treatment have been *self selected*. Compared to the rest of the class, they are eager beavers, and would probably have improved more than the rest of the class during this year, even without the tutoring, as a result of the numerous other efforts at self-improvement they would be making. What we need in the control group are pupils equally motivated, but lacking the tutoring.

Randomization in the Allocation of Scarce Resources

In settings where volunteering is an inevitable part of the setting, and in which the experimental treatment is a scarce resource in short supply, one useful strategy is to recruit volunteers in at least twice the number that can be accommodated. Again, as in the staged introduction, randomization can be presented as the fairest way of deciding who gets the treatment, producing an equivalent untreated control group as a by-product. Under such conditions, the volunteers understand that not all will get the treatment.

This device is admirable, although it creates in the control group an awareness of missed opportunity which could jeopardize external validity as a "reactive effect of arrangements." In the present study it has been decided to try to keep the control group from being aware of the availability of the treatment. Rather than calling for volunteers from the pupils, teachers will be asked to nominate matched pairs of pupils judged to need the assistance, to be likely to cooperate and to be residentially stable enough to be likely to remain in the school throughout the year. These pairs, matched also on sex and age, will be randomly split into experimental and control groups. At the end of the year, the teacher's records for these two will be abstracted to provide the basis of statistical comparisons. The experimental pupils will be individually contacted and invited to participate, with the tutor making every effort to put himself at the tutee's convenience.

Why do the instructions to teachers include estimates of residential stability and cooperativeness, when these requirements obviously bias the sample in the eager-beaver direction, reducing external validity and representativeness? The reason is an internal validity factor, and these must always have priority. If very few of the designated experimental group actually participate, this weakens or complicates the design considerably. The cooperating experimentals cannot be compared fairly with the controls as a whole, because they are a select subsample of the original experimental group. The most conservative treatment is to compare the total of the invited experimentals, participating plus non-participating, with the total controls. (Other approaches focusing upon the upper edge of gain scores for both groups, down to a cutting point determined by the number of participants, are being explored.)

The creation of a control group often goes against a therapeutic principle of giving scarce cures to the most needy. In spite of the democratic justification of randomization and the social gain from reality testing, many times this will take administrative guts. Let me give you the Saulk vaccine trials as a concrete illustration of this. They had a vaccine which they thought would save children from polio. If there were children that didn't get it, some of them were going to die "needlessly" and were going to be on their consciences. Think of the guts it took in this situation to give some children a saline solution in place of the Saulk vaccine, and to have those children dead next year, when if you had only given them the Saulk vaccine they would be alive. How did they get the moral fortitude to be able to do this? We need such fortitude, and we need the moral justifications that support such decisions. If the vaccine is in short supply, then if we distribute it on a random basis, we distribute it in such a way that we learn something from the limited supply that we have. We learn something that we couldn't learn if we had enough for everybody, nor if we used the procedure of giving first to those who needed it most (a procedure which too often also allows for the bias of special privilege to wealth). I am sure that in every school system you have many goods which are in too short supply to go around. With the democratic use of randomization, you can use these short supplies to create genuine experimental tests of the efficacy of that good.

Question: What happens to the tutors in this? From an educational point of view, we should be concerned with them and what happens to them as a result of their participation as well as what happens to the pupils they work with?

Comment: I have given you a feasible experimental design to measure the effect on the pupils that they are working with. This design is possible because we have more potential tutees than we have tutors. If the situation were reversed, and there were more volunteer tutors than tutees needing them, I could give you an equally elegant design. As it is, I have not been able to devise one, although I am fascinated by the effect on the attitudes of a sheltered upper-middle-class college student who has oversimplified the teachings of social psychology courses so as to learn not only that everybody is probably innately equal, but also they are in practice equal. (It is hard to connect the pupil's abilities with the vocabulary level of his home environment, the absence of previous opportunities, the fact that he has never had blocks and mechanical toys to play with. These connections are not "visible" while skin color is.) We could develop an annual attitude test and give it to all Northwestern students, and then pull out time-series analyses for those participating. But without a great deal of artificial apparatus, I see no easy experiment possible.

Strange as it may sound, one of the things I would advocate is opportunism. For the impact on pupils, an elegant laboratory is readily available. For the impact on the tutors, none is. Therefore, we study the pupils. Science, as it has existed, has been tremendously opportunistic, working first in the areas where it had laboratories. We should similarly leap to exploit the natural laboratories available in the social sciences. Colleges like Kansas State and Oberlin assign Negro and white students to dormitory roommates at random. Here is an impact situation that is interpretable, with excellent experimental and control groups available.

If the laboratory provided by the tutoring project proves effective, then we can opportunistically see what other problems can be tested in it. It will be particularly useful for comparing tutoring methods. Should you coach a child so that he never makes a mistake? Skinner and his teaching machines say, yes, allow no mistakes. The theory of knowledge lying behind the theory of experimentation I have been advocating (e.g., Campbell, 1960) if extended to the school situation, would say we have to learn by making mistakes. We have to give the pupil educational situations which will call out his predominant biases, so that he will learn they are wrong. Now these are quite different theories, and corresponding to them we could develop two ways of tutoring: one in which we only ask the child to make guesses where we have hedged the situation in with words so that he will guess right; the other of which we encourage him to expose many early guesses to reality testing.

Order of Experience for Interns: Counterbalanced Design

In one of the early discussions of this conference, Mrs. Evelyn Carlson gave as a minor but illustrative need for information the question as to whether the teaching interns being received from a group of non-urban colleges should be first assigned to a suburban school and then a central-city school, or whether the reverse order were better. She presented the question as one upon which social science advice should be useful. While in neither her mind nor mine is the problem of a level of importance comparable to the great bulk of those presented at this conference, it does provide an opportunity to illustrate some further points.

First, note the great range of generalization required from any analogous social science data. I think of two lines of evidence: Experiments by educational psychologists have shown that students get more problems right with an easy-to-difficult problem order than with a difficult-to-easy order. If slum schools are more difficult than suburban ones, then the suburb-urb order should leave the student with the greatest feeling of mastery. The other possibly relevant data comes from studies of primacy and recency in persuasion. These are of dubious extrapolability, and find *both* primacy and recency effects, usually canceling each other. In general, on matters relevant to Mrs. Carlson's question, social science knows little that is relevant, and has had so little opportunity to experiment, that if she finds out which is better, she should inform social science, not vice versa.

The experimental design which suggests itself immediately is what McCall in his classic *How to Experiment in Education* (1923) called the "rotation experiment," and today would be called a "cross-over design" or "counterbalanced design." The plan would be to split the interns into two groups, one getting the urb-surburb order, the other the suburb-urb. The design fits in well with housekeeping requirements, as the facilities in both areas are used at both periods. By noting from which group the greater number of teachers are recruited, etc., and by replicating the design for several years, an excellent experimental answer would be provided at almost no cost other than a few hours of bookkeeping and planning. Of course, one would probably want to supplement these institutional records with a brief opinion and career-plans survey.

Sources of Resistance to Experimentation and Evaluation

The steps required are accessible enough and have been advocated often enough so that administrative experimentation would be standard practice rather than almost unheard of were it not for important sources of resistance. In concluding this talk, it seems well to discuss some of these.

Costs: priority in the allocation of funds. For any administrator, there are financial demands and needs far in excess of available budgets, and this would continue to be so were the budget to be doubled or tripled. It is of the nature of social organizations that the budgetary demands for increased services, increased facilities and increased personnel in already existing functions are represented by persisting and vocal spokesmen for particular interests within the system. Evaluation and experimentation usually lack any self-interested sponsorship, have no natural pressure groups acting on their behalf. However, much lip service may be given to reality testing and experimentation, when it comes down to specific budgetary priorities, it always turns out that there are not enough funds to go around.

While administrative experimentation such as I have been talking about could be carried out with a 10 per cent increase in the burdens of existing record-keeping staffs and one day per month from one of the

existing record-keeping staffs and one day per month from one of the

existing executives, it may be well to organize record-keeping activities as a separate staff under the direction of a research-oriented professional, in a Research and Records Department, just to build into the system this personalized reminder of the importance of this function. Such a department could well afford to have in it a cost-accountant. In the long run, administrative experimentation is going to require the combined talents of social science (including educational research) and modern cost accounting (including "operations research"). Such departmentalization would provide an internal lobby to protect the funds for this important function.

Political Vulnerability from Knowing Outcomes

In many of the political climates in which we operate, hard facts evaluating organizational effectiveness are a political liability, while the mask of ignorance is a genuine protection. To return to an earlier example, Chicago's reform police superintendent Orlando Wilson was wise (albeit perhaps unintentionally) when he rendered the police records uncomparable with previous periods. (Although this may have been necessary, as his mutinous subordinates might easily have inflated the records just to embarrass him, easily done when a sizable portion of crimes have been going unreported.) Similarly, in these days many school systems are less vulnerable because their records and summary figures are "color blind," they in fact do not know which pupils or which teachers are Negro, nor in which schools Negroes are present. Records of over-all achievement, and trends over time, can be selectively cited by politicians for their political ends. One is often better off where standing can be argued rather than ascertained.

We can hope for less self-defeating political systems. Until they emerge, it may seem wise to restrict administrative experimentation and the attendant assembly of statistics to the comparison of alternatives realistically available to the administrator, to neither of which he is unalterably committed, or committed in a way in which the outcome could be embarrassing. If time series are used, these can be reported as departures from the secular trend, without publicizing what the secular trend is, etc. With a beginning of experimental experience on neutral issues, we can perhaps build a social system better able to tolerate reality testing. (In the Bureau of Labor Statistics figures on cost of living and on unemployment we are beginning to do this, though many an in-office official has wished at times that the data were not available.)

Although the machinery of administrative experimentation is available for use in the evaluation of single administrators, we should probably not use it for this purpose. Such results have no extrapolation value, cannot contribute to accumulative applied science. Similarly, over-all evaluations of total systems are both highly threatening and void of implications for policy decisions. It is better that we experiment with policy alternatives, where outcomes can provide generalizable bases for action.

No-Difference Outcomes

From more experienced research workers, dread of the very frequent outcome of no difference can serve as an inhibition. Such an outcome is painful and frustrating, and is worth avoiding for this reason alone. For longer and harder the experimental effort, the more torturing this pain. Furthermore, the experimenter may often feel that the lack of difference is due to the crudity of the measuring instruments--and institutional records are apt at such times to seem both crude and remote as symptoms from the expected effects. No-difference outcomes may be particularly expected where two policies each with strong advocacy are being compared--the advocacy gives evidence of probable (partial) truth to each.

Furthermore, the no-difference outcome when combined with the costs of changing policy, strongly reinforce the conservatives within the organization who favor a no-change policy. The potential administrative experimenter may correctly feel that change is in itself desirable as an organizational stimulant, even if the new policy is intrinsically no better than the old one, and may rationally prefer to introduce an innovation without evaluation rather than risk stalemate from a no-difference outcome. (If the value of chance per se were to become recognized, one might change the rules so that a new policy would always be introduced unless research showed it to be significantly worse than the old.)

Overconfidence in Decisions: Dissonance Reduction

The experimental approach connotes a tentativeness about administrative decisions, a willingness to consider the possibility that either alternative might be superior. This attitude is uncharacteristic of the decisive administrator, or, indeed, of anyone who is forced to make a total commitment to alternatives which may in fact be very evenly weighed, or ambiguous in their implications. Recently, Festinger (1957) and others (e.g., Brehm & Cohen, 1962) have provided evidence of a dynamic re-evaluation of alternatives which takes place after decision making, leading to a consolidation of the choice, creating a rigidity of decision. Such "dissonance reduction" devalues the rejected alternatives and enhances the chosen one. Such processes may be a necessary part of a decision-making process that is not going to oscillate in indecision. But its psychological concomitant is to create such an exaggerated faith in the decision made that evaluative research seems needless. Accompanying this is the dissonant fear that research might prove that alternative wrong.

The spelling out of these sources of resistance does not mean that I find them justifying inaction, nor compelling, even though each has its modicum of justification. Laying them out this way may give us the insight to overcome them.

SUMMARY

This talk has advocated administrative experimentation, using institutional records as measures of effect. Such experimentation could be a repetitive, continuous, in-service activity, an alternative to big one-shot research projects imposing outside staffs and measures upon the organization. The findings of such administrative research would have optimal generalizability to administrative decision making.

This orientation has then been applied to five concrete problems, illustrating an equal number of design alternatives and numerous analysis problems. Sources of resistance to administrative experimentation have also been discussed.

BIBLIOGRAPHY

Brehm, J. W. and Cohen, A. R. *Explorations in Cognitive Dissonance*. New York: Wiley, 1962

Campbell, D. T. Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin*, 1957, 54, 297-312.

Campbell, D. T. Methodological Suggestions from a Comparative Psychology of Knowledge Processes. *Inquiry*, 1959, 2, 152-182.

Campbell, D. T. Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity. *American Psychologist*, 1960, 15, 546-553. (a)

Campbell, D. T. Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes. *Psychological Review*, 1960, 67, 380-400. (b)

Campbell, D. T. From Description to Experimentation: Interpreting Trends as Quasi-Experiments. C. W. Harris (ed.), *Problems in Measuring Change*. Madison, Wisc.: University Wisconsin Press, in press.

Campbell, D. T. and Fiske, D. W. Convergent and Discriminant Validation by the Multitratit-Multimethod Matrix. *Psychological Bulletin*, 1959, 56, 81-105.

Campbell, D. T. and Stanley, J. C. "Experimental and Quasi-Experimental Designs for Research on Teaching." N. L. Gage (ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963.

Festinger, L. *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press, 1957.

McCall, W. A. *How to Experiment in Education*. New York: Macmillan, 1923.

Popper, K. R. *The Logic of Scientific Discovery*. New York: Basic Books, 1959.

Rivlin, H. H. *Teachers for the Schools in Our Big Cities*. October, 1962. (Mimeographed.)

OTHER MEASURES

A Survey of Unconventional and Cooperation-Free Measures for Social Science

by

Eugene J. Webb
Donald T. Campbell
Richard D. Schwartz

Supported in part by U.S. Office of Education Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

"Your bait of falsehood takes this corp of truth:

And thus do we of wisdom and of reach,

With windlasses and with assays of bias,

By indirection find direction out."

Hamlet, II, i

Preface

This monograph has had a series of working titles, and we should identify them for the benefit of our friends who shared early drafts. To some, this is "The Bullfighter's Beard"--a provocative, if noncommunicative title drawn from the observation that toreros' beards are longer on the day of the fight than on any other day. No one seems to know if the torero's beard really grows faster that day, because of anxiety, or if he simply stands further away from the blade, shaking razor in hand. Either way, there were not enough American aficionados to get the point, so we added ". . . and Other Nonreactive Measures."

This title lasted for a while, but the occasionally bizarre content of the material shifted the working title over to "Oddball Research," "Oddball Measures," and the like. Most of our friends have known the manuscript under one of the "oddball" labels, and it is only a fear of librarians that has caused us to drop it. In this day of explicit indexing, we feared that the book would nestle on a shelf between "Notes for the T Quarterback" and "Putting Hints for Beginning Golfers." As much as we might admire the company of a Ben Hogan, it is preferable outside a library.

Table of Contents

Chapter I 1

Chapter II Trace and Erosion: Physical Evidence 41

Chapter III Archives I: The Running Record 64

Chapter IV Archives II: The Episodic and Private Record 108

Chapter V Simple Observation 140

Chapter VI Contrived Observation: Hidden Hardware and Control 177

Chapter VII A Final Note 208

Chapter VIII A Statistician on Method 225

Chapter IX Cardinal Newman's Epitaph 226

Chapter 1

This survey directs attention to social science research data not obtained by interview or questionnaire. Some may think this exclusion doesn't leave much. It does. Our review of this material is intended to broaden the social scientist's currently narrow range of methodologies and to encourage creative and opportunistic exploitation of unique measurement possibilities.

Today, some 90% of social science research is based upon interviews and questionnaires. We lament this overdependence upon a single, fallible method. Among the reasonable objections to interviews and questionnaires are that they intrude as a foreign element into the social setting they would describe, they not only measure but also create attitudes, they elicit atypical roles and responses, they are limited to those who are accessible and will cooperate, and the responses obtained are produced in part by dimensions of individual differences irrelevant to the topic at hand.

But the principal objection is that they are used alone. No research method is without bias. Interviews and questionnaires must be supplemented by methods testing the same social science variables but having different methodological weaknesses.

In sampling the range of alternative approaches, we examine their weaknesses too. The flaws are serious and give insight into why we do depend so much upon the interview. But the issue is not choosing among individual methods. Rather it is the necessity for a multiple operationism, a collection of methods combined to avoid sharing the same weaknesses. The goal of this monograph is not to replace the interview but to supplement and cross-validate it with measures that do not require the cooperation of a respondent and which do not, themselves, contaminate the response.

Here are some samples of the kinds of methods we will be surveying in Chapters 2 to 6 of this monograph.

. . . The floor tiles around the hatching chick exhibit at Chicago's Museum of Science and Industry must be replaced every two weeks. Tiles in other parts of the museum need not be replaced for years. The selective erosion of tiles, indexed by the replacement rate, is a measure of the relative popularity of exhibits.

. . . The accretion rate is another measure. One investigator wanted to learn the level of whisky consumption in a town which was officially "dry." He did so by counting empty bottles in ashcans.

. . . The degree of fear induced by a ghost story telling session can be measured by noting the shrinking diameter of a circle of seated children.

. . . Chinese jade dealers of antiquity used the pupil dilation of their customers as a measure of the client's interest in particular pieces of stone.

. . . Library withdrawals were used to demonstrate the effect of the introduction of television into a community. Fiction titles dropped, nonfiction titles were unaffected.

Before getting into the details of such methods, we need to state a methodological framework within which both traditional and novel methods can be evaluated. The remainder of this chapter will be devoted to this.

Operationism and Multiple Operations

The social sciences are just emerging from a period in which the precision of carefully specified operations was confused with operationism by definition--a twist now increasingly recognized as an unworkable model for science. We wish to retain and augment the first aspect without committing the error of the second.

A consideration of the laws of physics as they are seen in the measuring instruments of physics demonstrates that no theoretical parameter is ever measured independently of other physical parameters, other physical laws. Thus, a typical galvanometer responds in its operations not only according to the laws of electricity but also to the laws of gravitation, inertia and friction. By reducing the mass of the galvanometer needle, by orienting the needle's motion at right angles to gravity, by setting the needle's axis in jeweled bearings, by counterweighting the needle point, and so on, the instrument designer attempts to minimize the most important of the physical forces irrelevant to his measurement purposes. As a result, the galvanometer reading may reflect, almost purely, the single parameter of voltage (or amperage, etc.). Yet from a theoretical point of view, the movement of the needle is always a complex product of many physical forces and laws. The impact of the irrelevant components varies depending upon the range and condition of values measured, as is well described by Wilson (1952).

Starting with this example from physics and the construction of meters, we can see that no meter ever perfectly measures a single theoretical parameter; all series of meter readings are imperfect operationalizations of the theoretical parameters they are intended to measure.

Truisms perhaps, yet they belie the mistaken concept of the "operational definitions" of theoretical constructs popularized in the social sciences. The inappropriateness is accentuated in the social sciences because there are no measuring devices as carefully compensated to control all irrelevancies as is the galvanometer. There simply are no devices designed with so perfect a knowledge of all the major relevant sources of variation. In physics, the instruments we think of as "definitional" reflect magnificently successful theoretical achievements and themselves embody classical experiments in their very operation. In the social sciences, our measures are still pre-theoretical. They tap multiple processes and sources of variance of which we are as yet unaware. At such a stage, the theoretical impurity and factorial complexity of every measure is not a nicety for idle technicians, but is overwhelmingly and centrally relevant in all measurement applications which involve inference and generalization.

The mistaken belief in operational definitions of theoretical terms has permitted social scientists a complacent and self-defeating dependence upon single methods--usually the interview or questionnaire. Exactly the opposite is the operational implication of the inevitable theoretical complexity of every measure: it calls for a multiple operationism, i.e., for multiple measures which are hypothesized to share in the theoretically relevant components but have different patterns of irrelevant components (e.g., Campbell & Fiske, 1959; Campbell, 1960; Garner,1954; Garner et al.,1956; Humphreys, 1962). Once a proposition has been confirmed by two or more independent measurement processes, the uncertainty of its interpretation is greatly reduced. Thus, it is through a triangulation of measurement processes that the most persuasive evidence comes. If a proposition can survive the onslaught of a series of imperfect measures, with all their irrelevant error, confidence should be placed in it.

Efforts in the social sciences at multiple confirmation often yield disappointingly inconsistent results. Awkward to write up and difficult to publish, such results confirm the gravity of the problem and the risk of false confidence that comes with dependence upon single methods (Kendall, 1963; Campbell & Fiske, 1959; Campbell & McCormack, 1957; Campbell, 1957).

Interpretable Comparisons and Plausible Rival Hypotheses

In this monograph we deal with methods of measurement appropriate to a wide range of social science studies. Some of these studies are comparisons of a single group or unit at two or more points in time; others compare several groups or units at one time, others purport to measure but a single unit at a single point in time and, to close the circle, some compare several groups at two or more points in time. For purposes of this discussion, we shall assume that the goal of the social scientist is always to achieve interpretable comparisons, and that the goal of methodology is to rule out those plausible rival hypotheses which make comparisons ambiguous and tentative.

Often it seems that absolute measurement is involved, and that a social instance is being described in its splendid isolation, not for comparative purposes. But a closer look shows that absolute, isolated measurement is meaningless. In all useful measurement, an implicit comparison exists when an explicit one is not visible. "Absolute" measurement is a convenient fiction and usually is nothing more than a shorthand summary in settings where plausible rival hypotheses are unimportant or so few, specific, and well-known as to be habitually taken into account. Thus, when we report a length "absolutely" in meters or feet we immediately imply comparisons with numerous familiar objects of known length, as well as implicit comparisons with a standard preserved in some Paris or Washington sanctuary.

Further, the development of such a facile lingua franca is dependent upon the existence of "solid" substances with which to build measuring rods. If the only measuring rods available were rubber, the plausible rival hypothesis that the measuring rod had changed length would rule out "absolute" measurement. If temperature, humidity, gravitational orientation, and barometric pressure all had marked effects on measuring rods, again the fiction of absolute measurement could never have evolved, and all scientific work would explicitly state transient comparisons interpreted as situationally specific. In detail, however, the physical measurement of length cannot be absolute, because there are no true solids from which to build measuring rods, and the length of a measuring rod is indeed subject to transient fluctuations. These facts are all obvious to the working experimental physicist. The idealization of certain measurement processes as definitional by the physicists N. R. Campbell (1923, 1938) and P. W. Bridgman (1927) however have provided social scientists with a very deceptive philosophical concept of measurement in which these practical facts have been neglected.

For the purposes of interpretable comparison in the social sciences there are probably no functionally "absolute" measurements at the present time. If you report to a sophisticated psychologist a person's IQ, he will immediately want to know what test, whether group or individual, whether verbal or "nonverbal," the age at which administered, the number of previous tests the person has had, whether the administrator was encouraging or discouraging; for all of these things and more have been found to affect the measured intelligence. Even such "fundamental" measurement operations as counting are rarely functionally absolute. For example, there are systematic errors of enumeration in national censuses which are different from time to time. Thus, earlier censuses in modern African states grossly understated population. The errors came not only because of missed families and residences, but also because parents believed it was bad luck or poor taste to brag about (report fully, i.e.) the number of children they had. In the most recent censuses, regional political leaders have been sensitive to the reality of national services allocated along population lines, thus leading to a competitive overcorrection. In some rural areas, for example, all persons were counted who "belonged there" by birth, even though they had moved to the city.

This inflation has been blunted a bit by the fear of constituents that they will be more heavily taxed if the true size of the family is known.

Whether or not the countervailing errors cancel each other out is dubious. The politicians are more likely to have their error prevail. But either way, the notion of absolute counting is suspect. For many other counts, ambiguity is present because of inconsistencies in the size and area of the units counted, as well as the population base and time span employed in the enumeration.

While the generalization may be somewhat too extreme, certainly the confidence with which a measure can be viewed as absolute contrasts enormously between social and physical measurement. The social scientist should examine each application of each measurement procedure anew for threats to valid interpretation, i.e., for plausible rival interpretations of his data. This attitude is the very opposite from the complacent definitional operationalism (Campbell, 1960; 1965, in press) which holds the measurement process as it occurs to be the preferred definition of a simple, singular, theoretical construct. Definitional operationalism has no way of acknowledging the known complexity and the systematic impurities and irrelevancies of each actual measurement process.

If measurement is regarded always as a comparison, there are three classes of approaches which have come to be used in achieving interpretable comparisons. First, and most satisfactory, is experimental design. Through deliberate randomization, the ceteris of the pious ceteris paribus prayer can be paribused. This may require randomization of the respondents treated, a randomization of occasions or a randomization of stimulus objects. In any event, the randomization strips of plausibility many of the otherwise available explanations of the difference in question. It is a sad truth that randomized experimental design is possible for only a portion of the settings in which social scientists make measurements and seek interpretable comparisons. The number of opportunities for its use may not be staggering, but where possible, experimental design should by all means be exploited.

A quite different and historically isolated tradition of comparison is that of index numbers. Here, sources of variance known to be irrelevant are controlled by compensating transformations of raw data. This is analogous to the compensated and counterbalanced meters of physical science which also control irrelevant sources of variance. The goal of this old and currently neglected social science tradition is to provide measures that are "absolutely" interpretable, that provide immediately meaningful comparisons across wide spans of time and social space. Real wages, intelligence quotients, and net reproductive rates are examples, but an effort in this direction is made even when a percentage, a per capita or an annual rate is computed (Webb, 1962a). Index numbers cannot be used uncritically because the imperfect knowledge of the laws invoked in any such measurement situation preclude computing any effective all-purpose measures. Furthermore, the use of complex compensated indices in the assurance that they measure what they were devised for has in many instances proved quite misleading. Often relationships established between the index and external variables are a product of but one of the components--not the compensated compound. Cronbach (1958) has described this problem well in his discussion of the accuracy of dyadic scores of interpersonal perception. In the older methodological literature the problem is raised under the term index correlations (e.g., Stouffer, 1934; Guilford, 1954; Campbell, 1955).

Despite these limitations, the problem of index numbers, which once loomed large in sociology and economics, deserves to be reactivated and integrated into modern social science methodology. The tradition is relevant in two ways for the problems of this monograph. Many of the measures that are suggested here will require a transformation of the raw data if they are to be interpretable in any but truly experimental situations. Such transformations should be performed with the wisdom accumulated within the older tradition, as well as with a regard to the precautionary literature just cited. Properly done, such transformations will often improve interpretability even if they fall far short of some ideal all purpose measurement that is uncritically applicable and functionally absolute (cf., Bernstein, 1935). In presenting specific measures, comments on possible indices will be made. Where no effective ones seem possible, the measure may in fact be usable only in true experimental conditions.

A second value of the literature on index numbers lies in an examination of the types of irrelevant variation which the index computation sought to exclude. The construction of index numbers is usually in response to criticisms of less sophisticated indices. They thus embody a summary of the often unrecorded criticisms of prior measures. In the criticisms and the corrections are clues to implicit or explicit plausible rival interpretations of differences, the viable threats to valid interpretation.

Take so simple a measure as an index of unemployment or of retail sales. The gross number of the unemployed or the gross total dollar level of sales is useless if one wants to make comparisons within a single year. Some of the objections to the gross figures are reflected in the seasonal corrections applied to both classes of time-series data. If we look at only the last quarter of the year, we can see that the effect of weather must be considered. Systematically the winter will depress the number of employed construction workers, for example, and increase the unemployment level. Less systematically, spells of bad weather will keep people in their homes and reduce the amount of retail shopping. Both periodic and aperiodic elements of the weather should be considered if one wants a more stable and interpretable measure of unemployment or sales. So, too, the presence of our gift-giving custom at Christmas spurs December sales, as does the coinciding custom of Christmas bonuses to employees. All of these are accounted for, grossly, by a conglomerate correction which is applied to the gross levels for either December of for the final quarter of the year.

Some of these sources of invalidity are too unique to a specific setting to be usefully generalizable; others are too obvious to be worth inventorying. But some will contribute to a general enumeration of recurrent threats to valid interpretation in social science measures.

The technical problems of index number construction are heroic. "The index number should give consistent results for different base periods and also with its counterpart price or quantity index. No reasonably simple formula satisfies both of these consistency requirements" (Ekelblad, 1962, p. 726). For most psychology and sociology applications, the major consistency issue is that of the base reversal test. If one is measuring at two points in time, A and B, with a raw score of 50 at time A and 100 at time B, which period is used as a base is a difficulty. If A is the base time, the score at B is twice as large, or 100 per cent greater, or has an index of 2.0. If Time B is the base, the score at A is only half as large, 50 per cent of the score or has an index of .5. One should be able to reverse bases and get a consistent index.

The time problems are resolved by substituting a geometric mean for an arithmetic one, but then other problems arise. With complex indices of many components, there is the issue of getting an index that will consistently yield scores across all the different levels and times of the components. In his important work on economic cycles, Hansen (1921) said, "Here is a heterogeneous group of statistical series all of which are related in a causal way, somehow or another, to the cycle of prosperity and depression"(p.21). The search for a metric to consistently relate these different components, to be able to reverse factors without chaos, makes index construction a difficult task. But the payoff is great, and the best approximation to solving both the base reversal and factor reversal issues is a weighted aggregate with time averaged weights. For good introductory statements of these and other index number issues, see Zeisel (1957) and Ekelblad (1962). More detailed treatments can be found in Mudgett (1951), Mills (1927) Fisher (1923) and Mitchell (1921).

The third general approach to comparison may be called that of "plausible rival hypotheses." It is the most general and least formal of the three. Given a comparison which a social scientist wishes to interpret, this approach asks what are the other plausible interpretations which the research setting and the measurement processes also allow. The more of these, and the more plausible each is, the less validly interpretable is the comparison. The solidarity of this approach depends upon the evolving criticism of social scientists. Their work provides the alternative interpretations appropriate to each case. However, it is also useful to employ a checklist of commonly relevant classes of threats to validity. The social scientist can go through such a list and judge the plausibility of each for his particular setting. Often many of them are irrelevant. And, for some major threats, it is often possible to provide supplementary analyses or to assemble additional data which can rule out a source of possible invalidity. The interpretation that comes from a review of alternative hypotheses is full of comparative considerations. Nonetheless it may become a valuable part of a cumulative social science, as repeatedly replicated findings have, in each instance supported different rival interpretations.

J. H. Wigmore, the distinguished legal scholar, wrote fifty years ago of the criteria of other plausible explanations of data:

> "If then, the potential defect of Inductive Proof is that the fact offered as the basis of the conclusion may be open to one or more other explanations or conclusions, the failure to exclude a single other rational hypothesis would be, from the standpoint of Proof, a fatal defect; and yet, if only that single other hypothesis were open, there might still be an extremely high degree of probability for the conclusion first claimed. . . . The provisional test, then, from the point of evidentiary fact point to the desired conclusion (not as the only rational hypothesis, but) as the hypothesis (or explanation) most plausible or most natural out of the various ones that are conceivable?" (Wigmore, 1913,pp. 18-19).

The notion in the law of "beyond a reasonable doubt" really means that no hypothesis is both present and plausible enough to rival the hypothesis of guilt.

While such an approach to a social science seems disappointingly hesitant, tentative, and tedious, there is still a sense in which it represents a bold disregard for unknown or unspecifiable weaknesses due to lack of control in a specific instance. For note that each plausible rival hypothesis is a threat only insofar as we give it the status of a law of equal or greater credibility as the law we seek to demonstrate. In some logical sense there are even in a "true" experimental comparison an infinite number of potential laws which singly or in complex combination could predict this result. We do not let this logical state of affairs prevent us from interpreting the results. Instead, uncertainty results only from those unexcluded hypotheses to which we, in the current state of our science, are willing to give the status of established laws: these are the _plausible_ rival hypotheses.

For those plausible rival hypotheses to which we give the status of laws, the conditions under which they would explain our obtained result also imply specific outcomes for other sets of data. Tests in other settings, attempting to verify these laws may enable us to rule them out. In a similar fashion, the theory we seek to test likewise has many implications other than that involved in the specific comparison, and the exploration of these is likewise demanded. The more numerous and complex the manifestations of the law, the fewer singular plausible rival hypotheses are available, and the more parsimony favors the law under study.

Our nostalgia is for data that prove and certify theory, but such is not to be our lot. Some comfort may come from the observation that this is not an existential predicament unique to social science. The replacement of Newtonian theory by relativity and quantum mechanics shows us that even the best of physical science experimentation probes theory rather than proves it. Modern philosophies of science as presented by Popper (1935; 1959; 1962), Quine (1953), Hanson (1958),Kuhn (1962), and Campbell (1965, in press) makes this point clear.

No knowledge is going to be certain, and social science should avail itself of those epistemological principles utilized in common sense knowing. These often result in satisfactory approximations of certainty of interpretation in uncontrolled situations. As illustrated elsewhere, the earthquake caused the brick building to crumble; the car crashing into it caused the telephone pole to break; the language patterns of the older models and mentors caused this child to speak English rather than Kwakiutl. Consider the inference as to a crashing car and the broken telephone pole. Our assurance of inference is due to the [practical] absence of other possible causes. Combinations of termites and wind we have ruled out because the other data implications of these theories (e.g., termite tunnels and debris in the wood, and wind records in nearby weather stations) do not occur. Spontaneous splintering of poles by happenstance coincident with the auto's onset does not impress us as too plausible nor does it help to explain the damage to the car (Campbell, 1963).

Internal and External Validity

Before presenting a checkoff list of some common sources of invalidity a distinction must be drawn between internal and external validity. Internal validity asks the question of whether a difference exists at all in any given comparison. It asks whether or not an apparent difference can be explained away as some measurement artifact. For true experiments, this question is usually not salient--but even there, the happy vagaries of random sample selection will occasionally delude one and spuriously produce the appearance of a difference where in fact none exists. For the rival hypothesis of chance we fortunately have an elaborated theoretical model which evaluates its plausibility. A *p*-value describes the darkness of the shadow of doubt which is never entirely absent. But for index number comparisons not imbedded in a formal experiment, and for the plausible rival hypothesis strategy more generally, the threats to internal validity--the argument that even the appearance of a difference is spurious, is a serious problem and the one that has first priority.

External validity is the problem of interpreting the difference, the problem of generalization. To what other populations, occasions, stimulus objects and measures may the obtained results be applied? The distinction between internal and external validity can be illustrated in two uses of randomization: when the experimentalist in agriculture or psychology randomly assigns a casual sample of plots or persons into two or more experimental groups, he is concerned entirely with internal validity--with making it implausible that the luck of the draw produced the resulting differences. When a sociologist carefully randomizes the selection of respondents so that his sample represents a larger population, representativeness or external validity is involved.

The first interview in a two stage study may set into motion attitude change and clarification processes which would otherwise not have occurred (e.g., Crespi, 1948). If that did occur, the comparison of a first and second measure on the same person is internally invalid, for the shift is a measurement-produced artifact—not one that would have occurred independently of measurement. Even when a measured control group is used, and a persuasive communication produces a greater change in an experimental group, then the persuasive effect may be internally valid, but externally invalid. There is the substantial risk that the effect occurs only with pretested populations, and might be absent in populations lacking the pretest (cf. Schanck & Goodman, 1939; Hovland et al., 1949; Solomon, 1949). For more extensive discussions of internal and external validity, see Campbell (1957) and Campbell and Stanley (1963).

The distinction between internal and external validity is often murky. In this work, we have considered the two classes of threat jointly, although occasionally detailing the risks separately. The reason for this is that the factors which are a risk for internal validity are often the same as those threatening external validity. A sociologist, for example, may be interested in representative sampling only to give him an internally valid comparison across cities or social groups. There is a muted line between the spurious effects of measurement itself and the effects of alternative social processes. Thus,we have taken an approach which contrasts with other efforts that have employed partially overlapping tests of risks to internal and external validity (Campbell, 1957; Campbell & Stanley, 1963).

Sources of Invalidity of Measures

In this section, 17 frequent threats to the interpretation of differences are presented. These are grouped under three general headings: Reactive Measurement Effects, Population Restrictions, and Content Restrictions. In presenting these, illustrations are provided, coming both from the verbal interview, questionnaire methods, and from the measures reviewed in this monograph. This section is the only one in which attention is paid to the dominating methods of current social science, and for that reason particular attention is paid to those weaknesses which create the need for multiple and alternate methods.

Reactive Measurement Effects

1. Respondent awareness of being measured, guinea-pig effect, self-consciousness, dissembling, and deceit face-saving.

Selltiz et al. (1959) make the observation:

> "The measurement process used in the experiment may itself affect the outcome. If people feel that they are 'guinea pigs' being experimented with, or if they feel that they are being 'tested' and must make a good impression, or if the method of data collection

> suggests responses or stimulates an interest the subject did not previously feel, the measuring process may distort the experimental results" (p. 97).

This effect was labelled "reactive arrangement" bias by Campbell (1957) and Campbell and Stanley (1963). It is important to note early that the awareness of testing need not, by itself, contaminate responses. It is a question of probabilities, and our feeling is that the probability of bias is high in any study in which a respondent is aware of his subject status.

Whereas the methods to be reviewed here do not involve "respondents," comparable effects on the population may often occur. Consider, for example, the potentially nonreactive instrument of the movie camera. If it is conspicuously placed, its lack of ability to talk to the subjects doesn't help us much. The visible presence of the camera no doubt changes behavior, and differentially depending upon the labelling involved. The response is likely to vary if the camera has printed on its side "Los Angeles Police Department" or"NBC" or "Harried Foundation Project on Unruly Crowd Behavior."

Similarly, an Englishman's presence at an African wedding exerts a much more reactive effect on the proceedings than it would on the Sussex Downs. A specific illustration may be of value. In the summer of 1952, graduate students in the social sciences at the University of Chicago were employed to observe the numbers of Negroes and whites in stores, restaurants, bars, theatres, etc. on a street (East 63rd) intersecting the Negro-white boundary. This, presumably, should have been a nonreactive process, particularly at the predominately white end of the street. No questions were asked, no persons stopped. Yet in spite of this hopefully inconspicuous activity, two merchants were agitated and persistent enough to place calls to the University which somehow got through to the investigators; how many others tried and failed cannot be known. The two calls were from a store operator and the manager of a currency exchange, both of whom wanted assurance that this was some University nosiness and not a professional casing for subsequent robbery. An intrusion conspicuous enough to arouse such an energetic reaction was also conspicuous enough to change behavior; had we been doing anything but counting noses, the bias would have been great. But even with the simple act of nose counting, there is the risk that the area would be differentially avoided. The error was in providing observers with clipboards and log sheets. Their appearance might have been still more sinister if they operated Veeder counters with hands jammed in pockets.

The present monograph also argues strongly for the use of archival records. Thinking, perhaps, of musty files of bound annual reports of some prior century, one might regard such a method as totally immune from reactivity effects. However, were one to make use of precinct police blotters, going around to copy off data once each month, the quality and nature of the records would almost certainly change. In actual fact, archives are kept indifferently, as a low priority task, by understaffed bureaucracies. Conscientiousness is often lax because of the lack of utilization. The appearance of a user can revitalize the process--as well as create anxieties over potentially damaging data, etc. (Campbell, 1968). The selective and wholesale rifling of records by ousted administrations sets an example of potential reactive effects, self-consciousness and dissembling on the part of archivists.

These reactive effects may threaten both internal and external validity, depending upon the conditions. If it seems plausible that the reactivity was equal in both measures of a comparison, then the threat is to external validity or generalizability, not to internal validity. If the reactive effect is plausibly differential then it may generate a pseudo-difference. Thus in a study (Campbell & McCormack, 1957) showing a reduction in authoritarian attitudes over the course of one year's military training, the fact that the initial testing was done in conjunction with an official testing program, while the subsequent testing was clearly under external university research auspices. This difference provides (French, 1955) a plausible reactive threat jeopardizing the conclusion that any reduction at all has taken place even for this one group, quite apart from the external validity problems of explanation and generalization. In many interview and questionnaire studies, increased or decreased rapport and increased awareness of the researcher's goals or decreased fear provide plausible alternative explanations of the apparent change recorded.

The common device of guaranteeing anonymity demonstrates concern for the reactive bias. Making public a response, or guaranteeing to hide one, will influence the nature of the response. This has been seen for persuasive communications, the validity of reports of brands purchased and the level of antisocial responses. There is a clear link between awareness of being tested and the biases associated with a tendency to respond to socially desirable items.

It may be that reactivity can be overcome by so structuring the task that the subject becomes insensitive to the testing situation. This may be pronounced in certain simulation apparatus. The student pilot knows that the Link Trainer isn't really a plane, but the fidelity of the simulation, as well as the intense commitment of attention demanded, may eliminate the testing contaminant.

This suggests that reactivity may be selectively troublesome within trials or tests of the experiment. Training trials may accommodate the subject to the task, but a practice effect may exist that either enhances or inhibits the reactive bias. Early responses may be contaminated, later ones not, or vice versa (Underwood, 1957).

Ultimately, the determination of reactive effect depends on validating studies of which very few examples are currently available. Behavior observed under nonreactive conditions must be compared with corresponding behavior in which various potentially reactive conditions are introduced. Where no difference in direction of relationship occurs, the reactivity factor could be discounted. As an example, the jury studies conducted by Strodtbeck and his associates uses a technique which is thought to reduce reactivity to a minimum. Jurors are drawn from the regular jury-pool and instructed by the judge to deliberate about the case in all seriousness. Potentially reactive effects stem from the fact that the "trial" is recorded and the mock jurors are aware that their decision will not have consequences for the parties. To test the reactivity effects of these conditions, one might record a number of actual cases and present them to mock juries, varying the method of selection of jurors, instructions to them, and other conditions. An intensive testing program of this kind might produce simulation technique whose results corresponded so highly to the decisions of the original juries that reactivity could be ruled out as a consideration.

In the absence of systematic data of this kind, we have little basis for determining what is and what is not reactive. Existing techniques consist of asking subjects in a posttest interview whether they were affected by the test, were aware of the deception in the experiment, and so forth. While these may sometimes invalidate a method as reactive they may be expected to miss many instances in which reactivity occurs. Subjects who consciously dissemble during an experiment may do so afterward for the same reasons. And those who are unaware of the effects on them at the time of the research may hardly be counted on for valid reports afterwards.

The types of measures surveyed in this monograph have a double importance in overcoming reactivity. In the absence of validation for verbal measures, nonreactive techniques of the kind surveyed here provide ways of avoiding the serious problems faced by more conventional techniques. Given the limiting properties of these "other measures," however, their greatest utility may inhere in their capacity to provide validation for the more conventional measures.

2. Role selection. Another way in which respondent's awareness of the research process produces differential reaction involves not so much inaccuracy, defense, or dishonesty, but rather a specialized selection from among the many "true" selves available in any respondent.

By singling out an individual to be tested (assuming that being tested is not a normal condition), the experimenter forces upon the subject a role defining decision--what kind of a person should I be as I answer these questions or do these tasks? In many of the "natural" situations to which the findings are generalized, the subject may not be forced to define his role relative to the behavior. For other situations, he will. In both cases, validity will be decreased as the role assumed in the research setting varies from the usual role present in comparable behavior beyond the research setting. The probability of this confounding role assumption, of course, varies from one research study to another. The novelty of a test-taking role may be particularly biasing for subjects with low education levels. Less familiar and comfortable with testing, those with little formal schooling are more likely to produce nonrepresentative behavior. The same sort of distortion risk occurs when the subject matter is novel. An unexpected subject matter may produce uncertainty in the respondent of which role to play. A role-playing choice is more conscious with such new or unexpected material.

This lack of familiarity with tests or with testing materials can influence response set in different ways. Responses may be depressed because of a lack of training with the materials. Or, the response level may be distorted as the subject perceives himself in the rare role of expert.

Both unfamiliarity or "expertness" can influence the character as well as the level of response. It is common to find experimental procedures which augment the experting bias. The instruction which reads, "You have been selected as part of a scientifically selected sample . . . it is important that you answer the questions . . ." underlines what a special situation and special person the respondent is. The empirical test of the experting hypothesis in field research is the extent of "don't know" replies. One should predict that a set of instructions stressing the importance of the respondent as a member of a "scientifically selected sample" will produce significantly fewer "don't knows" than an instruction set that does not stress this individual importance.

Although the "special person" set of instructions may increase participation in the project, and thus reduce some concern on the sampling level, it concurrently increases the risk of reactive bias. In science as everywhere else, one seldom gets something for nothing. The critical question for the researcher must be whether or not the resultant sampling gain offsets the risk of deviation from "true" responses produced by the experting role.

Not only does interviewing result in role selection, but the problem or its analogues may exist for any measure. Thus in a study utilizing conversation sampling with totally hidden microphone or with so stellar a mask as a mock hearing aid, each social setting will elicit a different role selection. Conversation samples might thus differ between two cities, not because of any true differences, but rather because of subtle differences in role elicitation of the differing settings employed.

3. Measurement as change agent. With all of the respondent candor possible, and with complete role-representativeness, there can still be an important class of reactive effects--those in which the initial measurement activity introduces real changes in what is being measured. The change is real enough in these instances, but is invalidly attributed to any of the intervening events, and is invalidly generalized to settings equivalent to the measurement process. This process has been deliberately demonstrated by Schanck and Goodman (1939) in a classic study involving information-test taking as a deliberately disguised persuasive process. Research by Roper cited by Crespi (1948) shows that the well-established "preamble effect" (Cantril, 1944) is not merely a technical flaw in determining the response to the question at hand, but also creates attitudes which persist and are measurable on subsequent unbiased questions. Crespi reports additional research of his own confirming that even for those who initially say "don't know," processes leading to opinion development are initiated.

The effect has been long established in the social sciences. In psychology, research in transfer of training early encountered the threat to internal validity called "practice effects" in which the exercise provided by the pretest accounted for the gain shown on the posttest. This research led to the introduction of control groups in studies that had earlier neglected to include them. Similarly, research in intelligence testing early showed dependable gains in test passing ability due to the experience with previous tests even where no knowledge of results had been provided. (See Anastasi, 1958, pp. 190-191, and Cane & Heim, 1950, for reviews of this literature.) Similar gains are shown in "adjustment" (Windle, 1954) although here increasing awareness of the examiner's purpose is more plausibly invoked.

While such effects are obviously limited to intrusive measurement methods, such as this review seeks to avoid, the possibility of analogous artifacts must be considered. Suppose that one was interested in measuring the weight of the women in a secretarial pool, and it was to be the dependent variable in a study of the effects of a change from an all female staff to one including men. One might for this purpose put free weight scales in the women's rest room, with an automatic recording device inside. However, the recurrent availability of knowledge of own weight in a semisocial situation would probably act as a greater change-agent for weight than would any experimental treatment that might be under investigation. Better would be a floor panel treadle recording weights without providing feedback to the participant, possibly disguised as an automatic door opener. An acute methodologist will see that this too is imperfect. A contaminant of the women wearing more and/or heavier costume jewelry is a risk not controlled by the treadle.

4. Response-sets. The critical literature in questionnaire methodology has demonstrated the presence of several irrelevant but lawful sources of variance. Most of these are probably applicable to interviews also although less elaborately demonstrated to date. Cronbach (1946) has summarized this literature, and continued evidence continues to show its importance (e.g., Chapman & Bock, 1958; Jackson & Messick, 1957).

Respondents, in general, more frequently will endorse a statement than they will disagree with its opposite (Sletto, 1938). This tendency differs widely and consistently among individuals, generating the reliable source of variance known as acquiescence response set. Another idiosyncracy, dependably demonstrated over varied multiple choice content, is the preference for strong statements versus moderate or indecisive ones. Sequences of questions asked in very similar format produce stereotyped responses--such as a tendency to endorse the right-hand, or the left-hand response, or to alternate in some simple fashion. Furthermore, decreasing attention produces reliable biases from the order of item presentation.

This can occur for questionnaires or public opinion polls [illegible], or for archival records such as votes (Bain & Hecock, 1957).

Suppose one wanted to obtain a nonreactive measure of the relative attractiveness of paintings in an art museum. He might employ an erosion method such as the relative degree of carpet or floor tile wear in front of each painting. Or, more elaborately, install invisible photoelectric timers and counters. This is an approach to be commended, but it must also take into account irrelevant habits which affect traffic flow. There is, for example, a general right turn bias upon entrance to a building or room. When this is combined with time deadlines and fatigue (do people drag their feet more before the paintings on the left side of the building?), there probably is a predictably biased response tendency. The design of museums tends to be systematic, and this too can bias the measures. The placement of an exit door will consistently bias the traffic flow and thus confound any erosion measure unless it is controlled. (For the imaginative and provocative studies on museum behavior see Robinson, 1928; Melton, 1933(a); 1933(b); 1935; 1936; Melton, Feldman and Mason, 1936.)

5. Interviewer effects. It is old news that the characteristics of the interview can contribute a substantial amount of variance to a set of findings. Interviewees respond differentially to visible cues provided by the interviewer. Within any single study, this variance can produce a spurious difference. The work of Katz (1942) and Cantril (1944) early demonstrated the differential effect of the race of the interviewer and that bias has been more recently shown by Athey, Coleman, Reitman and Tang (1960). Riesman and Ehrlich (1961) reported that the age of the interviewer produced a bias, with the number of "unacceptable" (to the experimenter) answers higher when questions were posed by younger interviewers. Religion of the interviewer is a possible contaminant (Robinson & Rohde, 1946; Hyman et al., 1954) as well as his social class (Riesman, 1956; Lenski & Leggett, 1960). Benney, Riesman and Star (1956) showed that one should think not only of main effects, but also of interactions. In their study of age and sex variables they report:

> "Male interviewers obtain fewer responses than female, and fewest of all from males, while female interviewers obtain their highest responses from men, except for young women talking to young men."

The evidence is overwhelming that substantial number of biases are introduced by the interviewer (See Hyman et al., 1954; Kahn & Cannell, 1957.) Some of the major biases, such as race, are easily controllable; other biases, such as the interaction of age and sex, are less easily handled. Even if we heeded all the known biases, without considering our ignorance of major interactions, there could no longer be a simple survey. The understandable action by most researchers has been to ignore these biases and to assume them away. The biases are lawful and consistent, and all research employing face to face interviewing or questionnaire administration is subject to them. Rather than flee by assumptions, there may be alternative methodologies that let one flee by circumvention.

6. Intra-instrument processes. The above threats to validity all represent processes within the respondent or social object of study. Another class of effects represent processes within the measuring instrument which might be confused with changes on the part of the persons or objects under study.

The measuring (data gathering) instrument is frequently an interviewer--whose characteristics, we have just shown, may alter responses. In panel studies, or those using the same interviewer at two or more points in time, it is essential to ask: To what degree is the interviewer or experimenter the same research instrument at all points of the research? Just as a spring scale fatigues with use, reading "heavier" a second time, an interviewer may also measure differently at different times. His skill may increase. He may be better able to establish rapport. He may have learned necessary vocabulary. He may loaf because of boredom. He may have increasingly strong expectations of what a respondent "means," and interpret more with practice. Note that some of these errors can be related to recording accuracy, while others are linked to the degree of the interviewer's interpretation of what transpires. There is always the risk that the interviewer is a variable filter over time and experience. Even where an increase in competence is involved, there is potential trouble. We usually think of difficulty only when the instrument decays. Yet a difference in competence between two waves of interviewing, either increasing or decreasing, can account for effects. Every measure should be questioned for this error source. Suicides in Prussia jumped 20 per cent between 1882 and 1883. This did not reflect a massive increase in depression so much as a change in record keeping. Until 1883 the records were kept by the police, but in that year the job was transferred over to the civil service (Halbwachs, 1930; cited in Selltiz et al.,1959). Chicago's recent crime rate increase seems due to new standards in record-keeping, not to a real change in incidence. Archivists undoubtedly drift in recording standards, with occasional administrative reforms in conscientiousness changing the calibration of the instrument (Kitsuse & Cicourel, 1963). Where human observers are used, they have fluctuating adaptation levels and response thresholds (Holmes, 1958; Campbell, 1961) sometimes induced by outcome orientation (Rosenthal, 1964).

Population Problems

7. Population restrictions. In the public opinion polling traditions one conceptualizes a "universe" from which a representative sample is drawn. This model gives little or no formal attention to the fact that only certain universes are possible for any given method. A method-respondent interaction exists--one that gives each method a different set of defining boundaries for its universe. One reason why so little attention is given to this fact is that, as methods go, public opinion polling is relatively unrestricted compared to other methods. Yet even here there is definite universe rigidity, definite restrictions on the size and character of the population able to be sampled. In the earliest days of polling, people were questioned in public places, probably excluding some 80 per cent of the total population. Shifting to in-home interviewing with quota-controls and no callbacks still excluded some 60 per cent--perhaps 5 per cent unaccessible in homes under any conditions, 25 per cent not at home, 25 per cent refusals, and 5 per cent through interviewers' reluctance to approach homes of extreme wealth or poverty and to avoid third floor walkups.

Under modern probability sampling with callbacks and household designation perhaps only 25 per cent of the population is excluded, 5 per cent as totally inaccessible through private residences (e.g., those institutionalized, hospitalized, homeless, transient, in the military, mentally incompetent, etc.), another 20 per cent through refusals, unavailability after three callbacks, and moved with no known address. This 20 per cent figure is that found for the model Elmira study in its first wave/Williams (1950), although other studies have reported much lower figures. Also to be considered in population restriction is the degree to which the accessible universe deviates in important parameters from the excluded population. This bias is probably minimal in probability sampling with callbacks, but great with catch-as-catch-can and quota samples. The deviation functions on the nature of the terminal sampling unit. Much survey research has centered on household behavior, and the great mass of probability samples employ a pre-listed household as the last unit. This frequently requires the enlistment of a household member as a reporter on the behavior of others. Since those who answer doorbells are overrepresented with the old and the young, and with women, this can be a confounding error.

When we come to more demanding verbal techniques, the universe rigidity is much greater. What proportion of the population is available for self-administered questionnaires? Payment for filling out the questionnaire will reduce the limitations a bit, but a money reward is selectively attractive--at least at the rates most researchers pay. A considerable proportion of the populace is functionally illiterate for personality and attitude tests developed on college populations. It is recommended to the reader that he read a set of test instructions, over the telephone, to a friend. The most highly predictable response is "Huh?"

Not only does task-demandingness create population restrictions, differential volunteering provides similar effects, interacting in a particularly biasing way when knowledge of the nature of the task is involved (Capra & Dittes, 1962).

The curious, the exhibitionistic and succorant are likely to well populate any sample of volunteers. How secure a base for generalization are volunteers when such groups are overrepresented while the shy, suspicious and inhibited are underrepresented? The only defensible position is a probability sample of the units to which the findings will be generalized. Even conscripting sophomores is better than relying on volunteers. It may be that a paper in the sociology of knowledge will someday be titled, "Anthrosophomorism in American Social Science."

Returning to the physical rigidity of sampling, what proportion of the total population is available for the studio test audiences used in advertising and television program evaluation? Perhaps 2 per cent. For mailed questionnaires the population available for addressing might be 95 per cent of the total in the U.S., but low-cost convenient mailing lists probably cover no more than 70 per cent of the families through auto registration and telephone directories. Those excluded are, again, highly selective. If, however, we consider the volunteering feature, where 10 per cent returns are typical, the effective population is a biased 7 per cent selection of the total. The nature of this bias, according to a recent study (Vincent, 1964) includes a skewing of the sample in favor of lower middle class individuals drawn from unusually stable "happy" families.

There are more television households in the U.S. than there are households with telephones. In any given city, one is likely to find more than

15 per cent of the households excluded in a telephone subscription list--and almost all of these will be at the bottom end of the socioeconomic scale. Among subscribers, as many as 15 per cent in some areas will not list their number, and an estimate of 5 per cent over-all is conservative. To indicate the nature of this bias, the three highest density areas for unlisted numbers are Westchester County, New York, Lake County, Illinois, and the Beverly Hills area of Los Angeles. The unlisted problem can be defeated by a system of random digit dialing, but this accelerates cost at least tenfold, and requires a prior study of the distribution of exchanges. Some 50 per cent of dialings are met with busy signals and "not-at-homes." Thus, for a survey without callbacks, the accessible population of 80 per cent (listed phone households) equals 40 per cent. If individuals are the unit of analysis, the effective sampling rate, without callbacks, may drop to 20 per cent. Random digit dialing will help; so, too will at least three callbacks, but both together may price a study beyond possibility.

Problems of this sort are even more acute for the research methods considered in the present monograph. While a few, such as sample solicitation, have the full population access of public opinion surveys, most have much more restricted populations. This inevitably creates problems of external validity, as well as one internal validity--discussed below under Temporal and Cross-Sectional Population Stability. Consider, for example, the sampling of natural conversations. What are the proportions of men and women whose conversations are accessible in public places and on public transport? What is the representativeness of social class, or role?

Joint utilization of data with nonidentical population boundaries. The population restrictions discussed here are apt to seem so severe as to traumatize the researcher and to lead to the abandonment of the method. This is particularly so for one approaching social science with the goal of complete description. Such trauma is, of course, furthest from our recommendations. While discussion of these restrictions is an appropriate background to their intelligent use and correction, there is need here for a parenthesis forestalling excess pessimism.

First, it can be noted that a theory predicting a change in civic opinion, due to an event A and occurring between two time periods, 1 and 2, might be such that this opinion shift can be predicted for many partially overlapping populations. One might predict changes on public opinion polls with that universe, changes in sampled conversation on commuter trains for a much smaller segment of it, changes in letters mailed to editors and the still more limited letters published by editors, changes in purchase rates of attitude-relevant books on the part of that minute universe, etc. In such an instance, the occurrence of the predicted shift on any one of these meters is confirmatory, and its absence discouraging. If the effect is found on only one measure, it probably reflects more on the method than on the theory. (E.g., Burwen & Campbell, 1957; Campbell and Fiske, 1959.) A more complicated theory might well predict differential shifts for different meters, and again, the evidence of each is relevant to the validity of the theory. The joint confirmation between pollings of high income populations and commuter train conversations is much more validating than either taken alone, just because of the difference between the methods in irrelevant components.

The "outcropping" model from geology may be used more generally. Any given theory has innumerable implications and makes innumerable predictions which are unaccessible to available measures at any given time. The testing of the theory can only be done at the available outcroppings, where theoretical predictions and available instrumentation meet. Any one such outcropping is equivocal, and all types available should be checked. The more remote or independent such checks, the more confirmatory their agreement.

Within this model, science opportunistically exploits the available points of observation. As long as nature abhorred a vacuum up to 33 feet of water, little research was feasible. When industrial skills made it possible to represent the same abhorrence by 16 centimeters of mercury in a glass tube, a whole new outcropping for the checking of theory was made available. The telescope in Galileo's hands, the microscope, the induction coil, the photographic emulsion of silver nitrate and the cloud chamber all represent partial new outcroppings available for the verification of theory. Even where several of these are relevant to the same theory, their mode of relevance is quite different and short of a complete overlap. Analogously, social science methods with individually restricted and nonidentical universes can provide collectively valuable outcroppings for the testing of theory.

The goal of complete description in science is particularly misleading when it is assumed that raw data provide complete description. Theory is necessarily abstract, for any given event is so complex that its complete description may demand many more theories than are actually brought to bear on it--or than are even known at any given stage of development. But theories are more complete descriptions than obtained data, since they describe processes and entities in their unobserved as well as in their observed states. The scintillation counter notes but a small and nonrepresentative segment of a meson's course. The visual data of an ordinary object are literally superficial. Perceiving an object as solid or vaporous, persistent or transient, involves theory going far beyond the data given. The raw data, observations, fieldnotes, tape recordings sound movies, of a social event are but transient superficial outcroppings of events and objects much more continuously and completely (even if abstractly) described in the social scientist's theory. Tycho Brahe and Kepler's observations provided Kepler with only small fragments of the orbit of Mars, for a biased and narrow sampling of times of day, days, and years. From these he constructed a complete description through theory. The fragments provided outcroppings sufficiently stubborn to force Kepler to reject his preferred theory. They were even sufficient to reject Newton's later theory had Einstein's better fitting theory then been available.

So if the restraints on validity sometimes seem demoralizing, they remain so only as long as one set of data, one type of method, is considered separately. Viewed in consort with other methods, matched against the available outcroppings for theory testing, there can be strength in converging weakness.

8. Descriptive access to population characteristics. In evaluating methods, their potential validity checks should be considered, as well as the differences in the universes they tap. What is contained in the data from alternative measures that gives descriptive access to that population for determining the specific nature of its restrictions? Internal evidence in early opinion polls showed their population biases when questions about prior voting and education showed marked discrepancies from previous election results and census data.

On this criterion, survey research methods again have great advantages. Observational procedures can check restrictions only on such gross and visible variables as sex, approximate age, and conspicuous ethnicity. Erosion methods such as relative wear on floor tiles offer no such intrinsic possibility. However, it is possible in many instances to introduce interview methods in conjunction with other methods for the purpose of ascertaining population characteristics. Thus commuter train passengers, window shoppers and waiting-room conversationalists can, on a sample of times of day, days of week, etc., be interviewed on background data, probably without creating any serious reactive effects for measures taken on other occasions.

9. Random sampling possibilities. In any procedure offering the joys and sorrows of excesses of data, unbiased or known-bias sampling procedures are a consideration. This is awkward and incomplete when sampling lists are unavailable, but practical and adequate methods have been developed in opinion survey research.

Novel methods should be examined for their possibilities in this regard. Often a time or space sample will offer a practical substitute for a person sample. For example, a study of the viewing of bus advertisements used a time stratified random triggering of an automatic camera (Politz, 1959). One could similarly take a photographic sample of bus passengers modulated by door entries as counted by a photo cell. A photo could be taken one minute after the entry of every 20th passenger. For some methods, such as the erosion methods, total population records are no more costly than partial ones. For some archives, temporal samples or agency samples are possible. For voting records, precincts may be sampled. But for any one method, the possibilities should be examined.

10. Temporal population stability. Just as internal validity is more important than external validity, so too is the stability of a population restriction more important than its absolute degree. Consider for example conversation sampling on a bus or streetcar. The population represented will differ between dry days and snowy days, between winter and spring, as well as by day of the week. These shifts would in many instances provide plausible

rival explanations of shifts in topics of conversation. Sampling from a much narrower universe, say conversational samples from an employees' restroom in an office building, would be preferable because the population is more stable over time. Comparisons of interview survey results over time periods are more troubled by population instability than is generally realized, because of seasonal layoffs in many fields of employment, plus status differentiated patterns of summer and winter vocations.

11. Cross-sectional population stability. Similarly, research populations available to a given method may vary from region to region, providing a more serious problem than is a population restriction common to both. Thus for a comparison of attitudes between New York and Los Angeles, conversation sampling in busses and commuter trains would tap such different segments of the communities as to be scarcely worth doing. Again, a comparison of employees washrooms in comparable office buildings would provide a more interpretable comparison. Through the advantage of background data to check on some dimensions of representativeness, public opinion surveys again have an advantage in this regard.

Content Problems

12. Content restrictions. Not only does each method have some rigidities with regard to populations tapped, it also has rigidities with regard to content covered. Observational methods can be used to yield an index of Negro-white amicability by computing the degree of "aggregation" or nonrandom clustering among mixed groups of Negroes and whites. This method could also be used to study male-female relations, or army-navy relations in wartime when uniforms are worn on liberty. But these indices of aggregation would be unavailable for Catholic-Protestant relations or for Jewish-Christian relations. Door-to-door

solicitation of funds for attitude-relevant causes is obviously plausible, but available for only a limited range of topics. For public opinion surveys, there are perhaps tabooed topics (though research on birth control and venereal disease has shown these to be fewer than might have been expected). More importantly, there are topics on which people are unable to report but which a social scientist can reliably observe.

Examples of this can be seen in the literature on verbal reinforcers in speech and in interviews. (For a review of this literature, see Krasner, 1958, as well as Hildum and Brown, 1956; Matarazzo, 1962a.) A graphic display of opportunistic exploitation of an "outcropping" was displayed recently by Matarazzo, Wiens, Saslow, Dunham and Voas (1964). They took tapes of astronaut to ground-communicator speech for two space flights and studied the duration of the ground-communicator's unit of speech to the astronauts. The data supported their expectations and confirmed findings from the laboratory. We are not sure if an orbital flight should be considered a "natural setting" or not, but certainly the astronaut and his colleagues were not overly sensitive to the duration of individual speech units. This observational literature has consistently produced findings on the effect of verbal reinforcers that could not be obtained by any direct questioning. The whole topic of face-saving deception cited under point 1 is relevant for consideration here also.

13. Dross-rate. In any given interview, a part of the conversation is irrelevant to the topic at hand. This proportion is the dross-rate. It is greater in open-ended, general, free-response interviewing than it is in structured ones with fixed answer categories; by the same token the latter are the more reactive. But in all such procedures, the great advantage is the interviewer's power to introduce and reintroduce certain topics. This ability allows a greater density of relevant data. At the other extreme is

unobserved conversation sampling, which is low-grade ore. If one elected to measure attitudes toward Russia by sampling conversations on public transportation, a major share of experimental effort could be spent in listening to comparisons of hairdressers or the Yankees' dominance of the American League. For a specific problem, conversation sampling provides low-grade ore. The price one must pay for this ore, in order to get a naturally occurring response, may be too high for the experimenter's resources.

14. Temporal content stability. The stability of content represented is another important consideration, generally well handled in public opinion surveys. Consider conversation sampling again: If one is attending to the amount of comment on race-relations, for example, the occurrence of extremely bad weather, or a president's assassination, may so completely dominate all conversations as to cause a meaningless drop. This is a typical problem for index making. In such an instance one would probably prefer some index such as the proportion of all race comments that were favorable.

15. Cross-sectional content stability. Where regional comparisons are being made, cross-sectional stability in the kinds of contents elicited by a given method is desirable.

Take the measurement of interservice rivalry as a research question. As suggested earlier, one could study the degree of mingling among men in uniform, or study the number of barroom fights among men dressed in different uniforms. To have a valid regional comparison, it must be assumed that the same incidence of men will wear uniforms in public places when at liberty. This is not true, partially because of past experience in a given area, partially because of the proximity to urban centers. If a cluster of military bases are close to a large city, only a selective group will wear uniforms off duty, and they are more likely to be the belligerent ones. Another comparison region may have the same level of behavior, but will be less visible.

The effect of peace will be to reduce the influence of the total level of the observed response, since mufti is more common. But if all the comparisons are in peacetime, it is not an issue. The problem occurs only if we elected to study the problem by a time-series design which cut across war and peace. To the foot on rail researcher, the number of outcroppings may vary because of war, but this is no necessary threat to internal validity.

Temporal Problems

16. Control over time and space sampling. Timing and the flexible sampling of occasions is an obvious desideratum, even though the theory of sampling of occasions is yet poorly developed.

Getting a complete list of appropriate occasions from which to sample is difficult frequently, and offers less promise than time sampling. Personnel evaluation studies have employed time sampling with considerable success. Observation during a random sample of a worker's laboring minutes efficiently does much to describe both the job and the worker (Thorndike, 1949; Ghiselli & Brown, 1955; Whisler & Harper, 1962).

Public opinion surveys have obvious limitations in this regard, which have led to the utilization of telephone interviews and built-in-dialing recorders for T.V. and radio audience surveys (Lucas & Britt, 1950; 1963). How else besides a recorder, could one get a reasonable estimate of the number of people who watch The Late Show?

Spatial Problems

17. Locational sampling control. Sampling of locations, such as bus routes, waiting rooms, shop windows, etc. needs to be developed as a science.

Obviously different methods present different opportunities and problems in this regard. Among the few studies which have seriously attempted this type of sampling the problem of enumerating the universe of such locations has proved extremely difficult (James, 1951). Location sampling has, of course, been practiced more systematically with pre-established enumerated units such as blocks, census tracts and incorporated places.

Any such enumeration of sources of invalidity is bound to be incomplete. Some threats will be too highly specific to a given setting and method to be generalized, just as will some opportunities for ingenious measurement and control. This list contains a long series of general threats that apply to a broad body of research method and content. It does not say that additional problems cannot be found.

So long as we maintain, as social scientists, an approach to comparisons that considers compensating error and converging corroboration from individually contaminated outcroppings, there is no cause for concern. It is only when we naively place faith in a single measure that the massive problems of social research vitiate the validity of our comparisons. We have argued strongly in this chapter for a conceptualization of method that demands multiple measurement of the same phenomenon or comparison. The overreliance on questionnaires and interviews is dangerous because it does not give us enough points in conceptual space to triangulate. We are urging the employment of novel, sometimes "oddball" methods to give those points in space. The chapters that follow illustrate some of these methods, their strengths and weaknesses and their promise for imaginative research.

Chapter 2

Trace and Erosion: Physical Evidence

The fog had probably just cleared. The singular Sherlock Holmes had been reunited with his old friend, Dr. Watson (after one of Watson's marriages), and both walked to Watson's newly acquired office. The practice was located in a duplex of two physician's suites both of which had been for sale. No doubt sucking on his calabash, Holmes summarily told Watson that he had made a wise choice in purchasing the practice that he did, rather than the one on the other side of the duplex. The data? The steps were more worn on Watson's side than on his competitor's.

In this chapter we will look at research methods geared to the study of physical traces surviving from past behavior. Physical evidence is probably the least used type of all the classes of data available to the social scientist, yet because of its ubiquity, holds flexible and broad gauged potential. It is reasonable to start a chapter on physical evidence by talking of Sherlock Holmes. He, and his paperbacked colleagues, could teach us much. Consider that the detective, like the social scientist, faces the task of inferring the nature of past behavior (who did the Lord of the Manor in?) by the careful generation and evaluation of current evidence. Some evidence he engineers (by questionning), some he observes (does the witness develop a tic?), some he develops from extant physical evidence (did the murderer leave his eyeglasses behind?). From the weighing of several different types of hopefully converging evidence, he will make a decision on the plausibility of several rival hypotheses. For example:

H_1: The butler did it.

H_2: It was the black sheep brother.

H_3: He really committed suicide.

This chapter discusses only the physical evidence, those pieces of data which were not specifically produced for the purpose of comparison and inference, but which are there to be opportunistically exploited by the alert. It should be emphasized that physical evidence has greatest utility in consort with other methodological approaches. Because there are easily visible population restrictions associated with physical evidence, they have not been employed so much as they should. It is difficult to fit a patently weak method into a single measure strategy based on definitional operationism. The less visibly weak questionnaire or interview looks to be more valid; and it may be if only one measurement approach is taken. In a multimethod strategy, however, one does not have to exclude data of any class or degree solely because it is weak. If the weaknesses are known and considered, the data are usable.

We could examine the relative popularity of different brands of cigarettes by noting the brands of the butts left in a randomly selected sample of ashtrays in public places. It might be less messy and more appropriate to employ tax receipts from the different manufacturers, employing available records. But tax receipts would not be a complete substitute measure for all purposes. Gross as they are, they could not tell us the preference distribution among certain social class groups. For this, we might go to a direct observational method and note what brands were smoked by predefined members of the groups. Or, we might go back, for corroborating evidence, to looking at the brands of butts left behind. Both the study of available records not prepared by the researcher and direct observation of outcroppings are closely linked to the study of physical evidence. Because they partially overlap with physical evidence, it is desirable to consider them together, thereby reducing the number of rival hypotheses sufficient to explain a comparison.

It may be helpful to discriminate between two broad classes of physical evidence, a discrimination similar to that between the cameo and the intaglio. On one hand, there are the erosion methods, where the degree of selective wear on some material yields the measure. Holmes' solution of the stairs on the duplex is an example. On the other hand, there are trace measures, where the research evidence is some deposit of materials. Immediately one thinks of anthropologists working with refuse piles and pottery shards. The trace measures could be further subdivided according to the number of units of evidence. We might have two subclasses: remnants, where there is only one or a few indicators of the past behavior, and accretions, where there is an accumulative body of evidence with more units, possibly deposited over a longer period of time. For purposes of simplicity, it is easier to consider just the two main divisions of erosion and trace.

Natural Erosion Measures

Let's look first at some erosion measures. Duncan (1963) was a member of a committee to set up a psychological exhibit at Chicago's Museum of Science and Industry. The committee learned that the vinyl tiles around the live chick hatching exhibit had to be replaced every six weeks or so; tiles in other areas of the museum went for years without replacement. A comparative study of the rate of tile replacement about the various museum exhibits could give a rough ordering of the popularity of the exhibits. Note that although erosion is the measure, the knowledge of the erosion rate comes from a check of the records of the museum's maintenance department.

In addition to this erosion measure, unobserved observation studies showed that people stand before the chick display longer than they stand before any of the other exhibits. With this additional piece of evidence, the question becomes whether or not the erosion is a simple result of people standing in one place and shuffling their feet, or whether it really does indicate a greater frequency of different people viewing the exhibit. Clearly an empirical question. The observation and the tile erosion are two partially overlapping measures, each of which can serve as a check on the other. The observation material is more textured for studies of current behavior, because it can consider both dimensions of numbers of viewers and duration of viewing. The erosion measure cannot simply measure the duration of viewing, but it does permit an analysis of popularity over time, and does so with economy and efficiency.

There are other rival hypotheses which should be considered for tile wear, as well as for any erosion study. In the case of the chick exhibit, one quite plausible hypothesis is that heat from the chick incubator causes the immediately surrounding tiles to be softer and thus more amenable to wear. This is a specific case of the general hypothesis of differential locational conditions producing a difference. Wind, temperature, precipitation and natural decomposition are all factors which could be differentially at work to yield a spurious difference.

Those readers who have attended American Psychological Association meetings have doubtless observed the popularity of conditioning exhibits displaying a live pigeon or monkey (a Skinner-boxed baby has also done well in recent years). This observation offers independent evidence for the general principle that dynamic exhibits draw more viewers than static ones. The hypothesis could be tested further by more careful comparing of tile wear about dynamic and static exhibits in the museum, making corrections for their nonrandom distribution. At least part of the correction would be drawn from the previously mentioned research by Melton (1936) on response sets systematically present in museum traffic flow.

The "natural" erosion of elevator buttons, or the area about them, could be another measure--one to index the frequency of stops on various floors. With this there is the selective sampling issue that only a subset of all elevators contain floor selection buttons, and this subset is probably not a representative sample of all elevators. Another bias might be the differential character of occupants on different floors. If a tearoom were on the third floor and a weightlifting gym on the fourth, an equivalent amount of erosion might be produced by a smaller number of weightlifters. This same bias would confound data coming from an attempted confirmation of the button erosion by noting floor tile erosion outside the elevators on each floor. With weightlifters heavier than teadrinkers their weight would bias this measure in exactly the same way that their greater energy in pushing the buttons biases the first. Here, the element of greater weight and strength gives so plausible an alternative hypothesis for both measures that they are not indicated as tandem methods.

The wear on library books, particularly on the corners where the page is turned, offers an example of a possible approach that does illustrate a useful overlap measure. One of the most direct and obvious ways to learn the circulation level of books is to check how many times each of a series of titles has been removed from a library. This is an excellent measure and uses the records already maintained for other purposes. But it is only an indirect measure for the investigator who wants to know the relative amount of reading a series of books get. They may be taken out, but not read. Indeed, there is a term in the publishing industry for titles of just this class: "non-books." (The expensive photo-essay collections are examples.) It is easy to establish whether or not there is a close relationship between degree of wear and degree of checkouts from the library. If this relationship is positive and high, then the hypothesis that books are taken out but selectively not read is accounted for. Note that the erosion measure also allows one to study the relative use of titles which are outside the span of the library withdrawal measure. Titles placed on reserve, for example, are typically not noted for individual use by library bookkeeping. This erosion measure might, however, be confounded by an anxiety factor. Certain books, being more anxiety provoking, would cause a great amount of sweating of the hands and hence would show more wear for the same amount of use.

An alternative accretion measure that could be used with the erosion measures is the amount of dirt accumulated on the book. This method would, of course, be valid only if the library did not employ regular bookdusters, and to the degree that wind currents and dust settling were not differential over the collection.

A variation of the erosion method has been suggested by Brown (1960) for studying the food intake of institutionalized patients--frequently a difficult task. If the question is one of over-all food consumption of some administrative unit (say a ward under special treatment conditions compared to a control ward), Brown makes the engagingly simple suggestion of weighing food trucks that enter and garbage trucks that leave. The unit can be varied to be an individual tray of food, the consignment to the floor or ward, or the total input and output of the hospital as a whole.

Natural Trace Measures

There are large numbers of useful natural remnants of past behavior that can be exploited. We can examine now a few examples of behavior traces which were laid down "naturally," without the intervention of the social scientist.

The detective story literature, again, is instructive. In a favorite example of the authors (Barzun, 1962), a case hinged on determining where a car came from. It was solved (naturally) by studying the frequencies to which the car's radio buttons were tuned. By triangulation of the frequencies, from a known population of commercial station frequencies, the source of the car was learned. Here was a remnant of past behavior (someone setting the buttons originally) that included several component elements collectively considered to reach a solution. Unimaginatively, most detective fiction considers much simpler and less elegant solutions--such as determining how fast a car was going by noting the degree to which insects are splattered on the windshield.

Lucas and Britt (1950) report a study which estimated an advertisement's readership level by analysing the number of different fingerprints on the page. The set of prints was a valid remnant and the analysis revealed a resourceful researcher. Compare this to the anthropologist's device of estimating the prior population of an archeological site by analysing data on the size of floor areas (Naroll, 1962). Among the consistently detectable elements in a site are good indicators of the floor area of residences. When these can be keyed to knowledge of the residential and familial patterns of the group, these partial data, these remnants, serve as excellent population predictors.

Sawyer (1961) recounts the problem of estimating liquor sales in Wellesley, Massachusetts. A city without any package stores, the usual devices of observation of purchase or study of sales records are of no help. Following the anthropological tradition of refuse study, Sawyer studied the trash carted from Wellesley homes and counted the number of empty liquor bottles.

The duration of the sampling period is a consideration in studying traces of any product in which consumption of a visible unit takes a long time. The study must cover a large enough span to guarantee that a trace of the behavior will appear if, in fact, it did occur. With whiskey consumption estimation, there is the further demand that account be taken of such possibly confounding elements as holidays, birthdays, sales in nearby retail outlets and unusual weather. This is particularly true if estimates are being made of the relative consumption of specific types of liquor. A heat wave will produce a substantial increase in the consumption of gin, vodka and rum, while depressing consumption of scotch, brandy and blended whiskeys. Depending upon the area, an unusually high level of entertaining will produce consumption of either more expensive or less expensive whiskey than is usual. The temporal stability of many common products that could be used to measure behavior is quite low.

Some accretion data are built up quickly, and the problem of deciding on the appropriate period of study is negligible. Take the debris accumulation of a "ticker-tape parade." The New York Sanitary Commission regularly notes how many tons of paper float down onto the streets during a parade. One might use this as material in estimating the enthusiasm of response for some popular hero. Because the Sanitary Commission has been reporting how hard they work for years, it is possible to employ a control level of tonnages showered down upon other heroes. Did Glenn get a more or less enthusiastic response from the New Yorkers of his day than did Lindbergh from the New Yorkers of his? At best these data are suggestive. The demise of the ticker-tape machine has meant a confounding of data for long term historical analysis. While at one time ticker-tape parades had a dominance of ticker tape in the air, today it is confetti. One can make corrections, of course, but they must be tenuous.

The accumulation of coins, not confetti, is another relatively rapid accretion measure. If one wanted to study the number of tourists in Rome, he could use as an auxiliary measure the number of coins plopped into the Trevi Fountain. This measure must be controlled for a time by nationality interaction. Americans, with their coin tossing response sets, are likely to comprise a higher share of the tourist population in the summer, while the more frugal French make up a larger share in the spring and fall. The coin to person estimation ratio would thereby vary seasonally. The plausibility a car came from. It was solved (naturally) by studying the frequencies to which the car's radio buttons were tuned. By triangulation of the frequencies, from a known population of commercial station frequencies, the source of the car was learned. Here was a remnant of past behavior (someone setting the buttons originally) that included several component elements collectively considered to reach a solution. Unimaginatively, most detective fiction considers much simpler and less elegant solutions--such as determining how fast a car was going by noting the degree to which insects are splattered on the windshield.

Lucas and Britt (1950) report a study which estimated an advertisement's readership level by analyzing the number of different fingerprints on the page. The set of prints was a valid remnant and the analysis revealed a resourceful researcher. Compare this to the anthropologist's device of estimating the prior population of an archeological site by analyzing data on the size of floor areas (Naroll, 1962). Among the consistently detectable elements in a site are good indicators of the floor area of residences. When these can be keyed to knowledge of the residential and familial patterns of the group, these partial data, these remnants, serve as excellent population predictors.

Sawyer (1961) recounts the problem of estimating liquor sales in Wellesley, Massachusetts. A city without any package stores, the usual devices of observation of purchase or study of sales records are of no help. Following the anthropological tradition of refuse study, Sawyer studied the trash carted from Wellesley homes and counted the number of empty liquor bottles.

The duration of the sampling period is a consideration in studying traces of any product in which consumption of a visible unit takes a long time. The study must cover a large enough span to guarantee that a trace of the behavior will appear if, in fact, it did occur. With whiskey consumption estimation, there is the further demand that account be taken of such possibly confounding elements as holidays, birthdays, sales in nearby retail outlets and unusual weather. This is particularly true if estimates are being made of the relative consumption of specific types of liquor. A heat wave will produce a substantial increase in the consumption of gin, vodka and rum, while depressing consumption of scotch, brandy and blended whiskeys. Depending upon the area, an unusually high level of entertaining will produce consumption of either more expensive or less expensive whiskey than is usual. The temporal stability of many common products that could be used to measure behavior is quite low.

Some accretion data are built up quickly, and the problem of deciding on the appropriate period of study is negligible. Take the debris accumulation of a "ticker-tape parade." The New York Sanitary Commission regularly notes how many tons of paper float down onto the streets during a parade. One might use this as material in estimating the enthusiasm of response for some popular hero. Because the Sanitary Commission has been reporting how hard they work for years, it is possible to employ a control level of tonnages showered down upon other heroes. Did Glenn get a more or less enthusiastic response from the New Yorkers of his day than did Lindbergh from the New Yorkers of his? At best these data are suggestive. The demise of the ticker-tape machine has meant a confounding of data for long term historical analysis. While at one time ticker-tape parades had a dominance of ticker tape in the air, today it is confetti. One can make corrections, of course, but they must be tenuous.

The accumulation of coins, not confetti, is another relatively rapid accretion measure. If one wanted to study the number of tourists in Rome, he could use as an auxiliary measure the number of coins plopped into the Trevi Fountain. This measure must be controlled for a time by nationality interaction. Americans, with their coin tossing response sets, are likely to comprise a higher share of the tourist population in the summer, while the more frugal French make up a larger share in the spring and fall. The coin to person estimation ratio would thereby vary seasonally. The plausibility of this can be partially checked by studying the nationality of the coins dropped in. Although most will be Italian lira, there may be enough foreign coins to provide a generalizable check.

Continuing the tourist research line, it can be noted that major U. S. ports of entry contain a conspicuous charity box for the collection of foreign coins. One could examine the national distribution of these coins to estimate the frequency distribution of the countries last visited while abroad. This particular sampling of coins is not a perfect measure, of course, and will have validity problems to the degree that people did not cash in their coins as they moved from country to country. This could be supporting evidence for an examination of the relative volume of traffic on airlines disembarking from various ports. The coin measure may be less valid, individually, than traffic figures, but is cheap, fast and provides another partially overlapping measure with which to make comparisons.

Experimenter Intervention

The methods discussed so far have all been ones in which the social scientist has taken the data as they come, and not intervened in any way to influence the frequency or character of the indicator material. There are conditions under which the social scientist can intervene in the data production process without destroying the nonreactive gains characteristic of trace and erosion data. He might want to do this, for example, to speed up the incidence of critical responses--a sometimes nagging annoyance with slowly eroding or accreting materials. Or, he might want to guarantee that the materials under study were in fact equivalent or equal before they were modified by the critical responses. The essential point is that his intervention should not impair the nonreactivity of the erosion and trace measures by permitting the subjects to become aware of his testing.

Controlled Erosion Measures

John Wallace, our former colleague, once noted that it would be possible to estimate the activity level of children by measuring the rate at which they wear out shoes. It is theoretically possible to start at any point in time with the shoes that children are wearing, measure the degree of wear and then later remeasure. The difference between the two scores might be a measure of the effect of some experimental variable. If the measurements were surreptcicious, the experimenter would be merely noting a naturally occurring event and not involving himself with the materials.

The approach is vulnerable to the problems implicit in starting with shoes of unequal degrees of wear. It is possible that the same unit of erosion is more quickly produced when shoes are older than when they are newer. Or, older shoes may be worn under a different set of occasions than newer ones, and the character of the abrasive surfaces may differentially vary across these occasions. To control for these variables, Wallace suggested that all children be supplied with a standard set of new shoes, thus washing out some of the unwanted variance. It would also be possible to speed up the measure by introducing shoes with soles and heels of less durable materials. Thus, a faster study should be possible with rubber soled tennis shoes than with leather topped and soled shoes. A reactivity issue to be faced is the awareness of the parents that their children's shoes are being measured. Even though the children may not be sensitive to being tested, the parents' awareness may unwittingly introduce some element of bias that confounds comparison. This error risk is similar to experimenter errors unconsciously admitted to data in scientific experiments (Rosenthal, 1962). It could be partially handled by using institutionalized populations, but the gain in internal validity is paid for by a loss in ability to generalize the findings.

Still another way to expedite the data gathering process is to unobtrusively coat an object with some substance that will record erosion. This may be performed to allow a test not otherwise feasible, or to permit a more valid record than the base material "naturally" permits. Waiting for elevator buttons to wear down, for example, might take months before a measurable change appears on reasonably priced equipment. It is possible to coat or cover the buttons with some substance (a clear soft plastic, e.g.) that erodes more quickly than the more durable plastic, aluminum or steel of the button itself.

Many college campuses contain a statue of some august figure whose nose is rubbed for luck at examination time. The American scene is dotted with such shiny nosed potentials for research on student anxiety. It is reasonable to expect the level of anxiety at any given examination period to be reflected in the number of nose rubbings. One way to obtain this measure would be to station an unobtrusive observer at the scene who records the rubbings. Moreover, he could report on certain visible characteristics of the nose-rubbers, such as sex, race and bearded or not. Equally possible is an approach which coated the nose with some properly eroding material--a material of known deterioration characteristics under the physical conditions to which the statue is exposed. This is a grosser measure than those obtainable through observation, but as in the case of the tile erosion method, offers an independent, overlapping check on the observational findings.

A careful investigator might want to improve upon the setting of the floor erosion approach. An important bias, not mentioned earlier, is that each footfall is not necessarily an independent event. Once a groove on a stair becomes visible, for example, those who walk on that stair are more likely to conform to the position of the groove than are those who walked before it became visible. This is partially due to the physical condition of the stair, which tends to slide the person's foot into the groove, and also possibly due to a response set tendency to follow in the footsteps of others. This may be partially controlled in newer settings by placing mats on the steps and noting their wear. The mats could also hide the already eroded grooves. From the puristic point of view, however, the only independent pressure could be that from the first step.

Controlled Trace Measures

Just as with the erosion measures, it is sometimes desirable for the researcher to tamper with the materials pertinent to a trace comparison. Noted earlier was a fingerprint study of advertising exposure (Lucas & Britt, 1950). Another procedure to test advertising exposure is the "glue-seal record" (Politz, 1957). For each pair of pages in a magazine, a small glue spot was placed close to the binding, and was made inconspicuous enough so that it was difficult to detect visually, or tactilly. The glue was of such a type that it would not readhere once the seal was broken. After the magazines had been read, exposure was determined by noting whether or not the seal was intact for each pair of pages, and a cumulative measure of advertising exposure was obtained by noting the total number of breaks in the sample issue. This method was developed because of the pervasive response set tendency of people to state, under questioning, that they had seen or read an advertisement when they had not in fact done so. This particular measure was valuable in establishing the degree to which there was a spurious inflation of recall of having seen or read an advertisement. It was not used alone, but in concert with more standard interviewing practices to provide a validity check.

The content restrictions of this method are substantial. It does not provide data for a single page or advertisement, but instead only indicates whether or not a pair of pages were exposed to the person's eye. There is no direct evidence as to whether or not the person even looked at advertisements which appeared on this pair of pages. It also does not indicate how many people may have been exposed to a given pair of pages. One or more openings yields the same response.

Finally, there is the possibility of accidental breakage of the seal--a contaminant which would overstate the true level of exposure. The Politz researchers recognized this hazard, and employed the measure in specifically controlled situations where the chance of breakage was almost zero. The fragility of the seal thus placed a great restriction on the method, for it had to be employed in settings where the subject had a high probability of being aware that he was being tested. This limitation thus added a population restriction because of the selective character of those who would come into a laboratory setting for testing, as well as the danger of the reactivity restriction which always accompanies laboratory research on humans.

The fingerprint method suffers from fewer restrictions, and it, too, could be improved by an unobtrusive move of the investigator. It is possible to select paper which more faithfully receives fingerprints than most of the paper used in normal magazine production, thereby reducing the risk that the level of exposure will be underestimated because of lack of a trace when a finger impression was made. The greater fidelity of a selected paper would also improve the ability to discriminate among the different fingerprints on the page. It is clearly impractical and unwise to base a complete study of advertising exposure on fingerprints; it is equally unwise to not consider coincidental methods which yield, as the glue seals do, independent validation data. Clearly, the greater the risk that awareness response set, role evocation and other variables present to valid comparisons, the greater the demand for independent, nonreactive, and coincidental measures.

From fingerprints to noseprints--and back to the museum for a final example. The relative popularity of exhibits with glass fronts or display cases could be compared by examining the number of nose prints deposited on the glass each day (or on some sample of time, day, month, etc.). This requires that the glass be dusted for noseprints each night and then wiped clean for the next day's viewers to smudge. The noseprint measure has less content restrictions than most of the trace techniques, for the age of viewers can be estimated as well as the total number of each exhibit. The age of viewers is determined by plotting a frequency distribution of the heights of the smudges from the floor, and relating these data to normative heights by age (minus, of course, the nose to top of head correction). There is an interesting contaminant in this measure. Because many fathers carry children on their shoulders, it could be possible to be misled into thinking that a substantial number of extremely tall people were coming to the museum. There is a correction that can be applied, of course, were one to find a pileup of responses in the area nine feet from the floor.[1]

Transforming the Data: Corrections and Index Numbers

The examples provided suggest that physical evidence data is best suited for measures of incidence, frequency, attendance and the like. There are exceptions. In a closely worked out theory, for example, the presence or absence of a trace could provide a critical test or comparison. But such critical tests are rare compared to the times when the physical evidence--be it deposit or erosion--is one part of a series of tests.

---

[1] One of us (RDS) feels that there might be an important response set bias with the noseprint measure: a "woodpecker" response pattern in which the same individual leaves multiple traces, thus resulting in an overestimate of frequencies. Although it is an empirical and not a theoretical question, DTC and EJW do not share their colleague's estimate of its plausibility.

When dealing with frequency data, particularly when they are in time-series form, it is essential to ask whether or not there are any corrections which may be applied to remove extraneous sources of variance. We can, of course, achieve more valid comparisons if known sources of error are removed.

More so than most classes of data, the type of frequency data widened by physical materials is subject to influences which can be known (and corrected for) without substantial marginal research effort.

The museum measures of noseprint deposits or tile erosion can serve as examples. We can use these data to answer questions about the popularity of a given exhibit over time. Are the hatching chicks as popular now as they once were? Is there a boom in viewing of giant panda exhibits? The answer to these questions might be of interest in themselves, or we might want them to evaluate the effect of some other variable. We could conceive of a study estimating the effect of newspaper stories on public behavior. Did a story on the birth of a baby leopard in the zoo increase the number of zoo visitors and the number of viewers of a leopard exhibit in the natural history museum? The museum data might be one more outcropping of an effect that could be tested and used in consort with other measures to evaluate an effect.

In looking at the museum data, we could consider the newspaper story question as a problem in the effect of an exogenous variable on a time series, and we are compelled to look for other sources of variance besides the story. We know something of the pool from which the critical responses come. Museum attendance varies seasonally (highest in summer, lowest in winter), varies cyclically (up on holidays, weekends and school vacations), and has had a strong secular movement upward over time. All of these are known influences on museum attendance which are independent of the newspaper story. They may be partialled out of the total variance of the time series or, in descriptive statistics, be controlled in index numbers. To the degree that these scores can be accounted for--in either inferential or descriptive terms--we achieve more sensitive research.

Auxiliary intelligence exists for most applications of physical evidence data. It may be contained in records kept for other purposes, or come from prior knowledge. Consider the problem of estimating advertising exposure by the glue seal method. In studying a single ad, our only measure is the proportion of pages on which seals are broken. This is desultory data unless it can be related to other information about advertising readership. With the question, "Is this a good or bad ad?" the data must be modified in some way to account for variance that comes from the characteristics of the medium carrying the ad. Thus, it is known that ads have lower exposure scores in magazines with a large number of advertising pages. So, too, scores will be lower in magazines which mass advertising pages together than in magazines which disperse the ads throughout the issue. Finally, readers of some magazines systematically and predictably read ads more so than do readers of other magazines. Each of these elements should be considered in evaluating the single ad under study. These extraneous factors might be combined into a baseline index which states the reasonable expectation for an ad appearing in a very thin issue of Magazine A, with a dispersion strategy of ad placement. The observed score for the ad under study can thus be transformed into a number which is in some way related to the expected value. This approach to "description" is the one argued in Chapter I--description and all research inferences are comparisons. When the control is not developed within the data gathering of the study (as it is not in most of the possible physical evidence measures) it must be generated by analysis of other available intelligence.

Administering such control opens the whole problem of the limitations of _ex post facto_ designs. For diverging discussion of this, see Selltiz, Jahoda, Deutsch and Cook (1959) and Campbell and Stanley (1963). Our position is that the weakness of the design is deleterious only to the extent that a single _ex post facto_ comparison is viewed as _the_ test of a research hypothesis. It may be possible to develop extremely convincing evidence using only _ex post facto_ reasoning. This could occur if one had a series of different outcroppings, each of which partially overlapped the other, but all of which pointed the same way.

With more elaborate comparisons than evaluating a single exhibit, the problems of extraneous variables and the need for their control in both inference and description becomes magnified. Keeping with the museum, let's take the problem of comparing two exhibits located in different sections of the museum. The same information used earlier--the known variation by season, holidays, etc.--is again core intelligence for comparison of the observed physical evidence. The concern becomes one of interactions. At those predictably high and low attendance times, we can expect a significant interaction between the accessibility of the exhibit and the over-all level of attendance. The interaction could bias both the number of viewers and the duration of time spent viewing. We should expect the more accessible exhibit to have a significantly higher marginal lead on noseprints during high attendance periods than during low attendance periods. Some of the interactive difference may come from the greater individual fatigue felt with large crowds, which might restrict the length of time each visitor spends in the museum. Or, the size of the crowds may slow movements so that a number of people with fixed time periods to spend in the museum don't get around as much. Or, there may be a population characteristic such that a larger share of those who visit museums during peak periods are indifferent viewers who lack a compelling interest in the over-all content of the museum. They may view either more casually or more erratically.

Corrections must be made for both the main effects and the interactions in such cases, and the easiest way to make them is to prepare corrections based on known population levels and response tendencies. We should speak of good luck nose rubbings per thousand students, or the rate of floor tile replacement in a specific display area per thousand summer visitors. These figures are first line transformations that are valuable. They can become more valuable if enough is known to consider them as the numerator of an index fraction. With the study of past behavior giving the denominator, indices can be produced which account for irrelevant variance and make for better comparisons.

Up to this point, we have been talking only of corrections applied to a single measure. Following in the index number tradition, we can also consider a piece of physical evidence either as one component of an over-all index composed of several classes of different classes of data or as an element in a set of physical evidence combined to make an aggregate index.

One might want an over-all measure, extending over time, of the completeness to which an institution is being used. This could supplement information on the extent to which it is being attended. In a library, for example, the various types of physical evidence available (dust collection, page wear,card catalogue wear) could be gathered together and each individual component weighted and then combined to produce the over-all index. Similarly for museums.

Assume that one wanted to note the effect of an anxiety producing set of messages on the alleged link between cigarettes and lung cancer. One could, in the traditional way, have a before-after questionnaire study which checked on both attitudes toward cigarettes and self reports of smoking. Or, one could observe. If the anxiety producing message were imbedded in a lecture, at some point toward the middle, it would be possible to observe the frequency of cigarettes lit during, before, and after the mention of the deleterious effects of smoking. Or, one could wire a sample of the chairs in the lecture hall and record the amount of squirming exhibited by the auditors at various points. (Happily, there has been an extensive development of squirm-registering apparatus.) Or, one could note the sale of books about smoking--ideally framing the setting so that equally attractive titles were available that argued the issue pro and con. Or, one could study the number of cigarette butts left on the floor in the lecture hall as a partial measure of short term effect, and the number of butts left in a sample of other settings (other lecture halls, dormitory rooms, etc.) as evidence of a more long term effect.

All of these alternatives are viable approaches to studying the effect of the message, with physical evidence an important element because of its ability to economically measure long term effects, and to extend the physical area of investigation beyond the immediate experimental setting. Depending on the degree of knowledge one had about the messages, the audience and past effects, it would be possible to construct an index with differential weights for the various component measures of effect. Each of the measures may be rapped for weakness, but taken cumulatively--as separate manifestations of a hypothesized effect--they offer greater hope for validity than any single measure, regardless of its popularity.

An Over-All Evaluation of Physical Evidence Data

The outstanding advantage of physical evidence data is its nonvisibility. The stuff of analysis is material which is generated without the producer's knowledge of its use by the investigators. Just as with employing available secondary records, one circumvents the problems of awareness of measurement, role selection, interviewer effects, and the bias that comes from the measurement itself taking on the role of a change agent. Thus, physical evidence is, for the most part, free of reactive measurement effects. It is still necessary to worry over possible response sets which influence the laying down of the data. With erosion measures, this might be so obvious a bias as the tendency for people to apply more pressure to stairs when going up than when going down, or the less obvious tendency to turn right.

With trace measures, there is the question of whether the materials have selectively survived or been selectively deposited. Do some objects have a higher probability of being discarded in public places than others? Or equally an issue, do some materials survive the intervening events of time better? Archeological research has always faced the problem of the selective survival of materials. Some of this selection comes from the physical characteristics of the material: clay survives; wood does not. Other selection comes from the potential value of the material which might be discarded. In writing of the small decorated stamps (seals) used by the ancient Mexican cultures, Enciso (1953) noted, "If any gold or silver was used, the stamps have yet to be found or have been melted long ago. Wood and bone have not survived the ravages of time. This may explain the abundant survival of clay stamps"(p. iv). In Naroll's (1956) phrase, there are "durable artifacts." We will discuss this bias also in the chapter on archival and available records, but it is significant here for the restrictions that it places on the content of physical evidence data and thereby the ability to generalize findings.

Any single class of physical evidence is likely to have a strong population restriction, and all physical evidence data is troubled by population problems in general. It is not, for example, easy to get descriptive access to the characteristics of a possible population restriction. One has the remnant of past behavior--a groove or a pile--and it says nothing by itself of those who produced the evidence.

We also must be cautious of physical data, for it may vary selectively over time or across different geographical areas. It is possible to get some checks on the character of these restrictions by employing supplementary methods such as the interview or the questionnaire. Some inferences about the character of populations bias may come also from a time-series analysis of the data, possibly linking the physical evidence data to time series of other data hypothesized to be selectively contributing to the variance. In a systematic investigation, a careful sampling of both times and locations is possible and internal comparison of the findings may offer some clues. But the assumption must be that any set of physical evidence is strongly subject to population restrictions, and supplementary information is always required.

Other methods also have population restrictions, and it may be possible to turn the fact of a restriction into an asset. There are, for example, certain subsets of the population who are virtually impossible to interview. In such situations, an enterprising investigator should ask whether the subject is leaving traces of behavior or material which will offer some help in inferring the subject's critical behavior. In this type of problem, the physical evidence is the supplementary data, and is used to fill in the population restrictions of other methods used concurrently.

As far as the content available to the reach of physical evidence methods, there are substantial limitations. It is not often that an investigator tests a theory so precise in its predictions that the appearance or absence of a single trace is a critical test of the theory. Most of the time, physical evidence is more appropriate for indexing the extent to which an activity has taken place--the number of footfalls, the number of empties tossed aside. Because these activities are influenced by many other variables, we seldom have an absolutely clean expression of some state of being--thus the necessity for corrections and transformations. Yet if enough information does exist, or can be produced, the content restrictions are controllable because they are knowable.

There is the positive gain that the amount of dross in physical evidence is low or negligible. Typically, what is measured is relatively uncontaminated by a body of other material which must be discarded as not pertinent to the research investigation. One can pinpoint the investigation closely enough to eliminate the dross--something not possible in the more amorphous method of conversation sampling, or of observation of "natural" behavior.

Compared to other classes of research methods, we have noted few examples of prior research using physical evidence. This is not through preference, but because we have been unable to find more. Physical evidence data is off the main stream for most psychological and sociological research. This is understandable, but still regrettable. The more visible weaknesses of physical evidence should preclude its use no more than should the less visible, but equally real, weakness of other methods. If physical evidence is used in consort with more traditional approaches, the population and content restrictions can be controlled, while providing a novel and fruitful avoidance of the errors that come from reactivity.

Chapter 3

Archives I: The Running Record

"Possibly a wife was more likely to get
an inscribed tablet if she died before
her husband than if she outlived him."

The tablet cited here is a tombstone, and the quote is from Durand's (1960) study of life expectation in Ancient Rome and its provinces. Tombstones are but one of a plethora of archives available for the adventurous researcher, and all social scientists should now and then offer up a touch of incense to those literate, record keeping societies which dutifully lay down so much material appropriate for novel analysis.

The purpose of this chapter is to examine and evaluate some uses of data periodically produced for some alien purpose, but which can be exploited by social scientists. These are the ongoing, continuing records of a society, and the potential source of variegated scientific data, particularly useful for longitudinal studies. The chapter that follows will look at more discontinuous archives, but here the data are the actuarial records, the votes, the city budgets, and the communications media which are periodically produced, and paid for, by someone other than the researcher.

Besides the low cost of acquiring a massive sum of pertinent data, the obvious gain of archival material is its nonreactivity. Whatever other errors may be found in the material, among them will rarely be a masking or sensitivity because the producer of the data knows that he is being studied by some social scientist. This gain, by itself, makes the use of archives attractive if one wants to compensate for the reactivity which riddles the interview and the questionnaire. The error risks implicit in archival research are substantial, as they are in any single class of method, but, to repeat our litany, if they are recognized and accounted for by multiple measurement techniques, they are not disturbing.

More so than other scholars, archeologists, anthropologists and historians have tangled with the problems of archival data. Obviously, they frequently have little choice but to use what is available and then to toy with corrections. Unlike the social scientist working with a contemporaneous problem, there is little chance to work still another method which will be pertinent to the problem, and which will circumvent the singular weakness of the records being employed.

Naroll (1962) recently reviewed the methodological issues of archives in his book, Data Quality Control. His central argument focuses on representative sampling. Does the archeologist with his 1000-year-old pottery shards, or the historian with a set of 200-year-old memoirs, really have a representative body of data from which to draw conclusions? It is one part of "Croce's Problem." Either one is uncertain of the data when only a limited body exists, or uncertain of the sample when so much exists that selection is necessary. Modern sampling methods obviate the second part of the problem. We can know, with a specified degree of possible error, the confidence we will place in a set of findings. But the first part of Croce's problem is not always solvable. Sometimes the running record is spotty, and we do not know if the missing parts can be adequately estimated by a study of the rest of the series. That is one issue. But even if the record is serially complete, the collection of the secondary sources impeccable,

5

and the analysis inspired, the validity of the conclusions must rest on assumptions of the adequacy of the original material.

There are at least two major sources of bias in archival records--selective deposit and selective survival. They are the same two concerns one meets in dealing with physical evidence data, the subject of chapter 2 Durand's study of the ancient Roman tombstones illustrates the selective deposit concern. Does a study of a properly selected sample of tombstones tell us about the longevity of the ancient Romans, or only of a subset of that civilization? Durand, as noted, suggests that the timing of a wife's death may determine the chance of her datum (CXVII-CLII) being included in his sample. Not only wives who die after their husbands may be underrepresented. There is, too, a possible economic or social class contaminant. Middle and upper class Romans were more likely to have a tombstone (and particularly one that survived until now) than those in the lower reaches of Roman society. This bias is a risk to validity only to the degree that mortality rates varied across economic or social classes--which they probably did. The more affluent were more likely to have access to physicians and drugs, which, given the state of medicine, may have either shortened or lengthened their lives. It is to Durand's credit that he carefully suggests potential biases in his data and properly interprets his findings within the framework of possible sampling error.

This same type of sampling error is possible when studying documents, whether letters to the editor or suicide notes. We know that a systematic response set bias exists among the editors. They try to present a "balanced" picture on controversial topics regardless of how unbalanced the mail. With the study of suicide notes, the question must be asked if suicides who don't write notes would have expressed the same type of thoughts had they taken pen in hand. Any inferences from suicide notes must be hedged by the realization that less than a quarter of all suicides write notes. Are both the writers and nonwriters drawn from the same population?

The demographer can't get new Romans to live and die; the psychologist cannot precipitate suicides. And therein is the central problem of historical data. New overlapping data cannot be obtained from the same or equivalent samples. The reduction of error must come from a close internal analysis which usually means fragmenting the data into subclasses and making cross checks. Another approach is feasible if one has reports on the same phenomenon by different observers, with different qualifications, at some past point in time (Naroll, 1960; 1961).

In examining an extinct culture, for example, one can compare the surviving reports made by those who lived among the people for a long period of time to those reports from casual visitors. Or, there can be a comparison of the reports from those who learned the indigenous language and those who did not. On those items on which there is consensus, there is a higher probability that the item reported is indeed valid. This consensus test is one solution to discovery of selective deposit or editing of material. It does not eliminate the risk that all surviving records are biased in the same selective way; what it does do is reduce the plausibility of such an objection. The greater the number of observers with different qualifications, the less plausible the hypothesis that the same systematic error exists. Sometimes selective editing creeps in through an administrative practice. Columbus kept two logs--one for himself and one for the crew. Record keepers may not keep two logs, but they may choose among alternative methods of recording or presenting the data. Sometimes this is innocent, sometimes it is to mask elements they gauge as deleterious. In economic records, bookkeeping practices may vary so much that close attention must be paid to which alternative record system was selected. The depreciation of physical equipment is an example.Often deliberate errors or record keeping policy can be detected by the sophisticate. At other times, the data are lost forever.

One more example may serve. A rich source of continuing data is the Congressional Record, that weighty but sometimes humorous document, which records the speeches and activities of the Congress. A congressman may deliver a vituperative speech which looks personally unflattering after some reflection. Since proofs are submitted to the congressman, he can easily alter the speech to eliminate his peccadillos. The naive research user of the Record may be deluded in an analysis of material which he thinks is spontaneous, but which is in fact studied.

A demurrer is entered. Even if the data were originally produced without any systematic bias that could threaten validity, the risk of their selective survival remains. It is no accident that archeologists are pottery experts. Baked clay is a "durable artifact" that cannot be digested and decays negligibly. Naroll (1956) comments that artifacts survive because they are not consumed in use, are indifferent to decay and are not incorporated into some other artifact so as to become unidentifiable. Discrete and durable, they remain as clues, but partial clues; other evidence was eaten, rotted, or re-employed.[1]

For the student of the present, as well as of the past, the selective destruction of records is a question. Particularly in the political area, the holes that exist in data series are suspect. Are they not there because survival of their contents would reflect in an untoward way on the administration? Have the files been rifled? If records are destroyed casually, as they often are during an office move, was there some biasing principle for the research comparison which determined what would be retained and what destroyed?

---

[1] Decay by itself is no problem. It only becomes one when the rate and distribution of decay is unknown. If known, it may become a profitable piece of evidence—as Libby's (1964) work with radiocarbon dating shows.

Estimating missing values in a statistical series, one is usually delighted if all but one or two values are present. This gives confidence when filling in the missing cells. If one or two holes are missing in a potentially politically significant series, one is less sanguine and more suspicious of his ability to estimate the missing data.

With so many of our records on paper, selective survival will become an even greater problem for future scholars. Common printing paper is very short lived, and 100 years of unprotected storage may render documents unusable. This is particularly true of newspapers, and historians have lamented the development of sulphur paper, which yellows, dries and flakes at a rapid rate. Aware of this decay, some newspapers are now microfilmed for permanent storage, but they are usually an atypical sample. They tend to be the "newspapers of record"--the New York Times, Le Monde, Die Welt, Pravda and so on. Who is microfilming the New York Mirror, and how many sets are available for the future student of those Americans who chose not to be concerned with wheat raising in Afghanistan? Material printed on high rag content paper and enameled surface paper will survive longer, but again, are these documents a representative research sample? The decline of parchment, like the development of the telephone, is a blow to students of historical records.

Actuarial Records

Birth, marriage, death. For each of these, societies maintain continuing records as normal procedure. Governments at various levels provide massive amounts of statistical data, ranging from the federal census to the town hall clerk's entry of Joan and Harry's charming wedding. Such formal records have frequently been used in descriptive studies, but they offer promise for hypothesis testing research as well.

Take Winston's (1932) research. He wanted to examine the preference for male offspring in upper class families. One could have interviewed prospective mothers in affluent homes, or fathers in waiting rooms. Indeed, one could pay obstetricians to ask, "What would you like me to order?" Other measures, nonreactive ones, might be studies of adoption records, the sales of different layette colors (cutting the data by the class level of the store) or the incidence of "other sex" names--such as Marion, Shirley, Jean, Jerry, Jo.

But Winston went to the enormous data bank of birth records and manipulated them adroitly. He simply noted the sex of each child in each birth order. A preference for males was indicated, he hypothesized, if the male/female ratio of the last child born in completed families was greater than that ratio for all children in the same families. With the detail present in birth records, he was able to segregate his upper class sample of parents by the peripheral data of occupation, etc. The same auxiliary data can be employed in any study to serve as a check on evident population restrictions--a decided plus for detailed archives.

This study also illustrates the time sampling problem. For the period studied, and because of the limitation to upper class families, his measure is probably not contaminated by economic limitations on the absolute number of children, a variable that may operate independently of any family sex preference. Had his study covered only the 1930's, or were he making a time-series comparison over decades with economically marginal families, the time factor could offer a substantial obstacle for valid comparison. The argument for the existence of such an economic variable would be supported if a study of the 1930's showed no sex difference among terminal children, but did show significant differences for children born in the 1940's.

Economic conditions are only one of the factors important for the criterion we earlier labelled "temporal problems." Wars, depressions and acts of God are all events which can pervasively influence the comparisons of social science data. The subjective probability of their influence may be tenuous, yet the ability to control that influence through index numbers and other data transformations is a reasonable and proper practice.

There are many demographic studies of fertility levels in different societies, but Middleton (1960) showed a shrewd understanding of archival sources in his fertility study. He developed two sets of data: fertility values expressed in magazine fiction and actuarial fertility levels at three different time periods. For 1916, 1936 and 1956, he estimated fertility values by noting the size of fictional families in eight American magazines. A comparison with the population data showed that shifts in the size of fictional families closely paralleled shifts in the true U.S. fertility level.

Middleton had a troublesome sampling problem. Since a small number of magazines continued publication over the period from 1916 to 1956, was the group of eight long-term survivors a proper sample? This durable group may not have been representative, but quite proper. The very fact that these eight survived the social changes of the 40 years argues that they probably reflected the society's values (or a sufficiently large segment of the society to keep the magazine economically alive) more adequately than those which failed. The issue was not one of getting a representative sample of all magazines, but instead of magazines which printed material that would have recorded more faithfully the pertinent research information.

Christensen (1960) made a cross-cultural study of marriage and birth records to estimate the incidence of premarital sex relations in different societies. He simply checked off the time interval between marriage and birth of the first child--a procedure which showed marked differences in premarital activity among cultures. His study illustrates some of the problems in cross-cultural study. The human gestation period is relatively constant across societies, and thus this was a more appropriate, and probably infinitely more valid, approach than hiring a series of informants to tell about the activities of their neighbors. Because of dietary and climatic differences, there will be some small differences among cultures in the mean gestation period, but this is a controllable element. Data on the incidence of premature births of second and later children for each society would permit the correction. A population problem to be guarded against in these cross-cultural studies, however, is the differential recording of births, marriages and the like. There are many societies in which a substantial share of marriages are not formally entered in a record keeping system, although the parties initially regard the alliance to be equally as binding as do those in other societies where records are more complete. The incidence in Mexico of "free-union" marriages is both extensive and selective--more centered in the working classes than in other groups (Lewis, 1961). The gestation period measure is a reliably constant indicator, but getting a broad sample base for using it may be selectively difficult across societies. The comparisons that come from such studies must be careful ones.

Simple marriage records alone were used by Burchinal and Kenkel (1962) and Burchinal and Chancellor (1962). The marriage records were used as a handy source by Burchinal and Kenkel to study the association between religious identification and occupational status. The records provided a great body of data from which to work, but also posed a sampling question. Are men about to be grooms a good base for estimating the link between religion and occupation? The small cadre of confirmed bachelors is excluded from the sample universe, and depending upon the dates of the records studied, there can be an interaction between history and groomdom.

A later study by Burchinal and Chancellor (1963) took the complete marriage and divorce records of the Iowa Division of Vital Statistics for the years 1953 and 1959. From these records, they compared marriages of same religion and mixed religion pairs for longevity. As might be expected, they found mixed marriages to be significantly shorter-lived than same religion ones. Of the mixed marriages, those in which Protestants described themselves as non-affiliated showed the highest divorce rate.

Of all the marriage record studies, probably none is more engaging than Galton's (1870) classic on hereditary genius. Galton used archival records to determine the eminence of subjects defined as "geniuses" and additional archives to note how their relatives fared on eminence. Few scientists have been so sensitive as Galton to possible error in drawing conclusions, and in a section on occupations, he notes that many of the judges he studied postponed marriage until they were elevated to the bench. Even so, their issuance of legitimate children was considerable. In Stein and Heinze's (1960) summary:

> "Galton points out that among English peers in general there is a preference for marrying heiresses, and these women have been peculiarly unprolific" (p. 87).

And on the possible contaminatn of the relative capacities of the male and female line to transmit ability:

> ". . . the decidedly smaller number of transmissions along the female line suggests either an 'inherent incapacity in the female line for transmitting the peculiar forms of ability we are now discussing,' or possibly 'the aunts, sisters and daughters of eminent men do not marry, on the average, so frequently as other women.' He believes there is some evidence for this latter explanation" (p.89).

Galton always had his eye turned to the plausibility of rival hypotheses.

Another pioneering study, Durkeim's Suicide (1951), shows an active exploration of archival source possibilities. He reached "The conclusion . . . that the social suicide rate can be explained only sociologically" (p. 299) by relating suicide levels to religion, season of the year, time of day, race, sex, education, marital status, doing all of this for different countries. All of these variables were obtained from available archives, and their systematic manipulation presaged the morass of cross-tabulations that were later to appear in sociological research.

Wechsler (1961) integrated three different classes of archival data in his correlational study of the relationships among suicide, depressive disorders and community growth. He went to the census for population change data, to mental illness diagnoses in hospital records, and to the vital statistics of the state to get the suicide incidence.

The final study we will cite on death records is Warner's (1959) work, The Living and the Dead. Death in Yankee City, and its accoutrements, were the subject of this multimethod research. Warner went to official cemetery documents to establish a history of the dead and added interviewing, observation and trace analysis as aids to his description of graveyards. "Their ground and burial lots were plotted and inventory was taken of the ownership of the various burial lots, and listings were made of the individuals and families buried in them" (p. 287).

His findings are of interest for what they say of response tendencies in the laying down of the physical evidence. Here the response tendencies, and the way in which they vary across social class groups, become the major clues to the analysis. Warner found the social structure of Yankee City mirrored (if this be the proper verb) in the cemetery; he found evidence on family organization, sex and age differentiation, and social mobility. For example, the father was most often buried in the center of the family plot, and male headstones were larger than female stones.[1] In some cases, Warner found that a family which had raised its social status moved the graves of their relatives from less prestigious cemeteries to more prestigious ones.

---

[1] This would be an interesting source of data for comparative analysis of different cultures. In matriarchal societies, for example, do you find the matriarch's stone substantially larger than the husband's? Does the husband get a marker at all? What are the differences in societies with extended _v._ nuclear family structures?

Warner's findings tie in with Durand's (1960) study of ancient Rome. In both studies, the relative dominance of the male was demonstrated by the characteristics of the tombstones.

A more recent commentary on tombstones comes from Crowald (1964), who wandered through Moscow's Novo-Devich cemetery, noting the comparative treatment of old czarists and modern communists. After noting that over Chekov's grave a cherry tree is appropriately blooming, he states:

> ". . . the cemetery also tells a quieter, more dramatic tale. Climbing out of some weedy grass is the washboard-sized marker of Maxim Litvinov, once a Stalin foreign minister and the wartime Soviet envoy to America. His mite of a marker reminds what happened to those who fell from Stalin's favor" (p. 12).

Just as in ancient Rome, the timing of a wife's death makes a difference in the nature of the tombstone. Here, this potential contaminant is used as a piece of evidence.

> "Novo-Devich does show, too, that things have changed in Russia since Stalin. For example, there is the great marble monument to Rosa Kaganovich. She was the wife of Lazar Kaganovich, the Stalin lieutenant booted from power in 1957 by Premier Nikita S. Khrushchev. Kaganovich is in full disgrace, but he fell after Stalin died. So his wife, who died in 1961, still got her big place in the cemetery. Fresh flowers decorate her marble" (p. 12).

These objects are just big and small pieces of stone to the uninformed, but to the investigator who possesses intelligence on those buried, and relates it to the stones, the humble and grandiose memorials are significant evidence.

Frequently, one has a choice among different archival sources, and directories are a useful alternative, whether of residents, association members, or locations. Ianni (1957) elected to use city directories as the prime data source in his study of residential mobility. An analysis of these directories over time allowed him to establish the rates of mobility, and then relate these mobility indices to the acculturation of ethnic groups.

It is obviously a tedious task to perform such an analysis, and the work includes a high amount of dross. If possible, such mobility levels might be more efficiently indexed by access to change of address forms in the post office. But the question here becomes one of population restriction. Is the gain in efficiency that comes from use of change of address forms worth the possible loss in completeness of sampling? The answer comes, of course, from a preliminary study evaluating the two sources of data for their selective characteristics. Either one is better than the more diffuse archives of the census, where the unit of analysis is the census tract and not the individual family.

For some studies, more selective directories are indicated, and the inclusion of a person in a directory serves as one element in the researcher's discriminations. Who's Who in America doesn't print everybody's name, nor does American Men of Science. Clark (1955) used both of these sources in his "study of some of the factors leading to achievement and creativity with special reference to religious skepticism and belief." Boring and Boring (1948) used American Men of Science to choose the psychologists studied in their useful article on the intellectual genealogy of American psychologists. Fry (1933) had earlier used Who's Who in America in a study titled "The religious affiliations of American leaders." (See also Lehman and Witty, 1931.)

Fry's work showed that if one depends on the editors of such directories for selective inclusion, one must also rely on the individuals listed for complete reporting. All of the problems associated with self report are present, for the individual has a choice of whether or not he will include all data, and whether he will report accurately. The archive serves as an inexpensive substitute for interviewing a large sample of subjects stratifed along some known or unknown set of variables. Fry found that a 1926 religious census showed 3.6 per cent of the general population to be Jews, while only .75 per cent of the entries in Who's Who in America were listed as Jews. Does this mean that Jews are less distinguished, discriminated against in being invited to appear in the directory, or is there selective reporting by the Jews of their religion? Fry gave a partial answer to the question by a check of another directory--Who's Who in American Jewry. He found 432 persons in this directory who had reported no Jewish affiliation in Who's Who in America. By raising the question of another plausible hypothesis for a comparison (3.6 per cent of the population compared to .75 per cent in the directory), he structured a question which was testable by recourse to another archival source of high pertinence for the hypothesis.

Babchuk and Bates (1962) went to the membership list of the American Sociological Association in their work, "Professor or producer: the two faces of academic man." Membership lists of professional organizations vary widely in their value. Given that the ASA membership list includes an efficiently obtainable list of sociologists, how good a sample is it of the profession? There are sociologists who are not ASA members, and they are likely to be qualitatively different from those who are members--particularly on the professor vs. producer dimension. And, in a niggling point, are sociologists a good sample of academic man?

Clark (1957) used the American Psychological Association's directory in his study of the psychological profession. For any study extending over a long time period, the APA directory can be frustrating. As the number of psychologists grew, the detail in the individual listing shrank. Thus, the number of items on which a complete time series could be produced were reduced as tighter and tighter editing took place. The measuring instrument was constant in its content for only a few pieces of information. The instrument's change in its categories of information is a detectable shift in the nature of the measure. Other changes, such as increasingly difficult requirements for membership or individuals responding to the greater bulk of the directory by writing more truncated listings, may change the character of the instrument in a less visible way, and produce significant differences which are only recording artifacts.

Digging into the past, Marsh (1961) obtained the names of 1047 Chinese government officials from the government directories of 1778 and 1831 - 1879. He then correlated the ranks of the officials with the time required to reach a particular rank and with other factors such as age and family background. If there were no differential recording, one may conclude with Marsh that the rich get there faster.

Political and Judicial Records

An archival record--votes--is the dependent variable for officeholders; it is the absolute criterion measure. Votes cast by the people determine the politician's most important piece of behavior (staying in office) and votes cast by the legislator are the definitive test of his position and alliances.

Dozens of studies are available that have evaluated voting statistics by party or by individual. The ad hoc rhetorical condemnations of an opponent's record are at one end of a scale that is anchored by sophisticated factor analysis on the other. In the interests of space, we have limited our examples to studies which have reported either more unique or more convoluted manipulations of the essentially simple data.

The political slant of legislators has been a popular research topic. "Progressivism" in the U.S. Senate was assessed by Gage and Shimberg (1949). They wanted a sample of votes which would measure progressivism and picked ten bills for one congress and eight for another--getting coefficients of reproductability of .88 and .91 respectively. With these data they studied a series of questions: are younger senators more progressive than older ones? (no); do senators from the same state tend to vote the same? (no); are regional differences significant? (yes).

MacRae (1954b) also studied the same type of legislative tendency in his study of the influence of constituent pressure and congressional social grouping influence. He selected his sample of critical votes by consulting the roll calls published by the New Republic and the CIO News. His assumption was that these sources would only publish reports of votes on issues germane to their presumably liberal readers. From these then, he obtained a "liberal index" depending upon the direction of the vote.

A finer break was made by Dempsey (1962), who estimated conservative votes, but divided them along party and non-party issues. His subject was party loyalty in the Senate and he used the reports of roll call votes on individual senators to provide a "Loyalty Shift Index."

Moving over to the House, and the employment of all roll call votes, Riker and Niemi (1962) looked at the question of congressional coalitions. They took votes on 87 roll calls, noting whether a congressman (1) voted on the winning side,(2) voted on the losing side,(3) did not vote when eligible, did not vote when ineligible. The rolls were classified into subsets and finally an index of coalitions was produced.

Farris (1958) also used roll call votes to study coalitions with the Congress. His article is of strong interest, because it details the methodological issues the political scientist faces in isolating ideological groups. Farris elected to use Guttman scaling techniques on a sample of roll call votes from the House in the 79th Congress. His sampling of votes was studied and he excluded 94 of the 231 possible roll calls because: either there was no quorum, the vote was nearly unanimous, or sharp partisanship was shown. His scales included bills on "foreign policy" and "labor." ". . . it is possible to construct three-, four-, and five-position ideological groupings by cross-tabulating members' positions on the several analytic issues" (p. 328).

An elaborate set of analyses conducted by MacRae (1954a) illustrates a more complex analysis of multiple archival sources. MacRae asked how politicians with different seniority reduced the inherent insecurity of their elective jobs. The results are too extensive to report, but these are among the variables used:

seniority (number of consecutive terms of office)
number of representatives from each party elected in different years
rates of vote getting performance (an index of the legislator's vote compared to a control of the gubernatorial vote by the candidate of the same party)
primary election performance (ratio to nearest competitor)
Guttman liberal-conservative scale of voting issues
voting behavior on key bills

Only two classes of data are used--general election statistics and legislative voting--but MacRae's analysis indicate the richness of these archives for the venturesome student.

The political scientist must work with roll call votes; the desultory ayes and nayes on voice votes are never traceable to their sources. An empirical question (one on which there is no substantial research study) is the difference between bills voted on by roll call or by voice vote. It is reasonable to expect that a difference exists, and that some systematic criterion is at work. It may be that only the more significant votes get a roll call, but also possible is that some significant bills go through on a voice vote because leaders of both parties choose to avoid the record on some sensitive issues.

Taking the roll calls with proper hedging, though, there is the choice of which bills yield the best evidence for a particular research question, as well as directional decisions such as, "what is the liberal stand on this bill?" The population of congressmen voting sometimes varies substantially also, for some decide to evade a vote and a public stand on the bill. The "pairing" system, in which an absent congressmen may announce what his vote would have been had he been present, eliminates some of this absentee error, but so long as congressmen avoid a vote, some population restriction is present.

The content analysis of political speeches is another workaday practice of politicians and diplomats. In a study of group tension, for example, Grace and Tandy (1957) studied 13 speeches made by Soviet delegates before the League of Nations. Political tension may be indexed in many ways (cf. Bugental, 1948) but one method is to search for archival evidence of activities designed to reduce tension and uncertainty.

For congressmen, one device is to study the degree to which they make use of their franking (free mail) privilege. The rate of mail sent from congressional offices varies systematically, in a pattern closely linked to the proximity of the election year. We are not aware of any such study, but it should be possible to get an indirect measure of a congressman's job security by evaluating the use of this privilege. There are many speculations, for example, on what defines a "safe" seat for reelection to the congress. A journalistic benchmark is a classification based on the extent of the congressman's victory. If he were elected by more than 55 per cent of the vote, his seat is described as safe; if by less than 55 per cent, it is described as dangerous. It would be of interest to study the extent of mail sent (correcting for different size constituencies) by the margin of victory in the last election. Such an analysis, combined with other intelligence, may provide a more empirical definition of how congressmen themselves define as "safe" and "dangerous" seats.

The behavior of the congressmen can also be used to study those outside the congress. It is a common practice for a congressman to insert into the Congressional Record columns which reflect his point of view. In a study of political columnists, Webb (1962b ; 1963) employed these data for an estimate of liberalism-conservatism among twelve Washington columnists. Individual members of congress were assigned a liberal-conservative score by evaluations of their voting record published by two opposing groups--the conservative Americans for Constitutional Action and the liberal Committee on Political Action of the AFL-CIO. The two evaluations correlated -.75 for Senators and -.86 for House members. Columnists were then ordered on the mean score of Congressmen placing their articles into the Record.

In another section of the same study, a time-series analysis was made of three veteran columnists (Walter Lippmann, Arthur Krock and David Lawrence) and the data showed that a variable of the party affiliation of the inserter (defined not as Republican or Democrat, but as a member of the "in" or "out" party at the time of insertion) produced significant differences. A possible flaw in such a measure is that the inclination of the endorser (inserter) might not parallel that of the endorsee (the columnist). One could imagine a speech headed by, "Why _even_ David Lawrence said today. . . ." This possible contaminant was checked and found not to be a significant issue in the insertion into the _Record_.

It would be valuable to supplement such an analysis with a comparative study of how various writers treat the same event. Cogley (1963) suggests this in his poem on the interpretation of the Papal encyclical, _Pacem in Terra_:

> David Lawrence read it right
> Lippmann saw a Liberal light,
> William Buckley sounded coolish,
> Pearson's line was mostly foolish. . .
> To play the game you choose your snippet
> Of Peace on Earth and boldly clip it.
> (p. 12)

In off year and primary elections, only a small share of eligible voters cast ballots. But this selective behavior is not damaging to validity since the critical variables--election or not, margin of victories--are posited directly on this selectivity. Lustig (1962) studied pro-integrationist voting by matching aggregate data on demographic characteristics and pro-integrationist votes. Precinct votes in a southern campaign between a segregationist and an integrationist were compared to census information. This is an admittedly gross approach, but the sanctity of the voting booth precludes, usually, a direct study of individual voting behavior. One can go back to survey questionning, and relate these findings to the actual voting records, but the error in self-report of voting behavior is so great that such a potentially reactive approach is highly questionable. Digman and Tuttle (1961) have provided one of the few pieces of research in which investigators were able to randomly sample individual ballots. For most archival studies, however, this degree of precision is obtainable.

The voting records of the people can also be used to measure the effect of experimenter intervention in pre-election settings. Among the memorable studies in social science is Gosnell's (1927) field experiment on getting out the votes. Selecting twelve electoral precincts, Gosnell divided them into experimental and control conditions. The experimental precincts were sent a series of nonpartisan messages encouraging registration and voting in a forthcoming elections while the control precincts received only the normal pre-election stimulation--generally of a politically partisan nature. The effect of the mail effort was determined by an analysis of registration lists, pollbooks and census material. Here Gosnell intervened in a "natural" way, established controls, and employed inexpensive archival records as his tellingly appropriate dependent variable.

Bain and Hecock (1957) also demonstrated the ability to test persuasion principles in a natural laboratory free from the reactive biases of the university cubbyhole. They were interested in the effect of physical position on alternative choices, and found data in the aggregate voting statistics from Michigan elections. Michigan was chosen because of (1) the absence of a law requiring ballots to be burned after the election (an obvious impediment to archival studies of voting behavior), and (2) the systematic rotation of candidate names on the Michigan ballot.[1] Because of the Michigan system of rotation, an orthogonal analysis was possible, establishing the vote of each candidate and each position on the ballot. The findings supported the assumptions of veteran political hands and the ballot constructors: there was a significant position effect. It would be a provocative study to take this and other naturally occurring possibilities for a position effect (e.g., the sale of goods in a supermarket when placed at different shelf positions), and compare the results with those derived from the traditional experimental laboratory. For a field study, this one using "car cards," the advertising material in buses and subways, see Bureau of Applied Social Research (1953). Star and Hughes (1950) conducted a field propaganda study--evaluating the effect of a massive pro-United Nations campaign in Cincinnati.

---

[1] This rotation is practiced in several states, and represents an assumption that position on the ballot does indeed have an effect. "Under (California) state law, the incumbent's name goes first on the ballot, and political handicappers give as much as a 20 per cent edge--greater than the margin of most senatroial victories--to this psychological primacy". (Anonymous, 1964, 28).

Schwartz and Skolnick (1962) proposed a study of positive and negative incentives in tax compliance using changes in taxpaying in experimental and control groups as the dependent variable measure. This study would depend on the cooperation of the Internal Revenue Service which cannot legally disclose information concerning the returns of any individual taxpayer (Schwartz, 1961). The problem can be avoided by group comparisons in which the individual's identity is not revealed to the researcher. Another device is to study tax compliance in a state such as Wisconsin where individual returns are legally available for research purposes.

The votes of a judicial body also are data for other than obvious research. Kort (1957; 1958), Schubert (1959, 1963) Nagel (1962) and Ulmer (1963) have employed mathematical analyses of past voting behavior by U.S. Supreme Court justices to predict future votes--a more systematic attempt at the common game played by working lawyers and constitutional law experts.

The same body of information, U.S. Supreme Court decisions, was used by Snyder (1959) in a study of the degree of uncertainty in the total U.S. judicial system. In one measure, uncertainty was defined by the number of reversals of lower court decisions by the high court. With the precedent principle of stare decisis at work in our court system, there should be few reversals if certainty is high. Moreover, to the degree that there is certainty (predictability) of outcome upon appeal, there will be cases fought up through the inferior courts.

There is a very real population restriction in using such data--one that has been differential over time. It is a highly plausible hypothesis that appelate cases brought before the U.S. Supreme Court are not representative of the body of cases appearing before the inferior courts. Historically, a substantial share of high court cases have been those which have involved affluent litigants, for the cost of steering a case through the courts is demanding of both money and time. Some of the change in representativeness over time has come from the increasing affluence of our society, but more has come from the growth of well-financed vested interest groups who will assume the costs of litigation for a party. The spurting growth within civil rights groups of both legal talent and money documented by Vose (1959) and Krislov (1963) has meant an increasing number of cases before the appelate courts of cases that might have never been appealed fifty years ago. Thus, any comparison over time of the behavior of the court system relative to some legal issue must take account of this variation.

Green (1961) demonstrated the large store of judicial data available to the social scientist. He gathered a sample of 1,437 cases from 1956-1957 police and court records of Philadelphia, in order to study uniformity in sentencing and the criteria by which sentences were decided. Three sets of variables were isolated as sentencing criteria: legal factors, legally irrelevant factors and factors in the criminal prosecution. The relative severity of different types of sentences was measured by the extent of deprivation of civil liberty, and he compared how sentences differed according to variations in the criteria for the sentences. A series of nonparametric tests "provide assurance that the deliberations of the sentencing judge are not at the mercy of his passions or prejudices but comply with the mandate of the law."

In later work, Snyder (1959) expanded her uncertainty research to include indiviudal justices of the U.S. Supreme Court. Her measure was the frequency of dissenting opinions. Unlike the congress, if a vote is taken by the court, each individual justice participates in the decision.

There may be some bias from the selection of cases (a content restriction), for only a portion of the cases submitted to the Supreme Court are granted certiorari, i.e., accepted for ruling. In research on some issues, it may be that the court has systematically excluded from consideration pertinent cases, or has excluded a critical subclass. As in all cases of archival analysis, it is necessary to examine whether or not confounding research population restrictions exist. The advantage of legislative and judicial records is that one can learn something of the nature of these restrictions. It is a matter of record which legislators did not vote on roll calls, and which cases were rejected for consideration by the court.

Other Government Records

Some government records are orthodox sources for the social scientist. The birth, death and marriage archives, as we have already seen, can be used for straightforward descriptive work or for less direct applications. Other records have less visible, but equally fruitful, applications. In this section are examples of research which has used power failures, municipal water pressure, parking meter collections and the like as research data.

The weather is a resonable start. Durkheim (1951), as noted above, used weather as one of the variables in his study of suicide. An early investigation by Lombroso (1891) also used archival analysis to note the effect of weather and time of the year on scientific creativity. He drew a sample of 52 physical, chemical and mathematical discoveries and noted the time of their uncovering.

His evidence, shaky as it is, showed 22 of the major discoveries occurred in the spring, 15 in the autumn, 10 in summer and five in winter.

Mindak (1963) took the ongoing records of parking meter collections as one of his indices of the effect of a Minneapolis newspaper strike. He hypothesized that one of the major effects of the strike would be a decrease in retail shopping. Since most of the parking meters were located in the downtown shopping area, revenue collections from them was a good piece of evidence on the strike's effect. The data showed marked decreases during the months of the strike, using a control of previous years. This control accounted for some of the seasonal variation that might be present in the time series, but it did not account for the possibility that the decrease was part of a long term secular movement in the series. Had data been gathered for a period of months after the settlement of the strike, the plausibility of this hypothesis could have been tested by a before, during and after analysis.

City budgets were the stuff of Angell's (1951) study on the moral integration of American cities. He prepared a "welfare effort index" by computing local per capita expenditures for welfare and combined this with a "crime index" based on FBI data to get an "integration index."

A particularly and novel use of data comes from the study of city water pressure as it relates to television viewing. For some time after the advent of television, there were anecdotal remarks about a new periodicity in water pressure levels--a periodicity linked to the beginning and end of programs. As the television show ended, so the reports ran, a city's water pressure dropped as drinks were obtained and toilets flushed. A graphic display of this hypothesis provided by Mabley (1963), who published a chart showing the water pressure for the city of Chicago on January 1, 1963. This was the day of a particularly tense Rose Bowl football game, and the chart shows a vacillating plateau until the time the game ended when the pressure level plummets downward.

Using this approach, one could study the relative popularity of different retirement times. Since a large amount of water is used by many people at bedtime, a comparison of the troughs at 9:00, 9:30, 10:00 and so on could be made. Similarly a comparison could be made in the morning hours to estimate time of arising. Two problems arise. Do those who retire early use the same amount of water as those who retire late? It is, after all, possible that a smaller number of showers and baths would be taken by those who retire late--particularly in areas with a high share of apartment dwellers. Another difficulty of such a study is comparison across different time areas. Many people close out the day by viewing television news. In the metropolitan Chicago area for example, some three million people watch the 10:00 p.m. newscasts. But the last major television newscast in the Eastern areas is at 11:00 p.m. This one hour variation might influence the times of people going to bed. The water pressure index could help to establish whether it did or not. Similarly, it could be used to study the relative amount of attendance paid to entertainment and commercial content of television. The critical point of study in this research would be the water pressure levels at the times of mid-show commercials. The prior decision of viewers to turn the set off at a specified hour would influence the water pressure index for commercials at the beginning and end of shows, but should be a minor rival hypothesis for those embedded within the entertainment content of the program.

A similar measure, a more catastrophic one, has been manifest in the United Kingdom. This measure, electric power failures, gives plausibility to the hypothesis that it is commercials (and not earlier decisions toturn the set off) which determine the drops in water pressure. At the time of the introduction of the United Kingdom's commercial television channel, a series of power failures hit the island. Whole areas were blacked out, and it was noticed that the timing of the power failures coincided with the time of commercials on the new television channel. The explanation provided was that viewers left their sets to turn on electric water heaters to make tea. The resulting power surge, from so many heaters plugged in simultaneously, overloaded the capacity of a national power system unequipped to handle such peaks. The commercials remain; new power stations have been built. This measure is a more discontinuous one than water pressure, but it would be of value to compare the water pressure levels with power demands. If the hypothesis is correct that the English were plugging in water heaters, one should find a higher correlation between the two measures in the United Kingdom (with a small time lag as water power) than in the United States.

Another adept link between two time series of archival data was provided by DeCharms and Moeller (1962). They gathered the number of patents issued by the U.S. Patent Office from 1800 to 1950. Relating these to population figures, they prepared a patent production index for 20 year periods over the 150 year span. These data were then matched to findings from a content analysis of children's readers for the same period, with a prime focus on achievement imagery. The matching showed a strong relationship between the amount of achievement imagery in their sample of books and the number of patents per million population.

The Mass Media

Among the most easily available and massive sources of continuing secondary data are the mass media. The variety, texture and scope of this enormous data pool has been neglected for too long. In this section, we present a selected series of studies which show intelligent manipulation of the mass media. We have excluded the morass of descriptive content analyses and focused on that research which uses the media as a mirror of society, as a fecund recording source, or as a subject by itself for socially significant research.

It is proper to start this section by citing Zipf, who sought order in diverse social phenomena by his inventive use of data that few others would perceive as germane to scientific inquiry. In a model study, Zipf (1946) looked at the determinants of the circulation of information. His hypothesis was that the probability of message transfer between one person and another is inversely proportional to the distance between them. (See also Miller, 1947; Stewart, 1947; Zipf, 1949.) Without prejudice for content, he made use of the content of the mass media, as well as sales performance. How many and how long were out of town obituaries in the New York Times? How many out of town items appeared in the Chicago Tribune? Where did they originate? What was the sales level in cities besides New York and Chicago of the Times and Tribune? To this information from and about the mass media, Zipf added other archival sources. He asked the number of tons of goods moved by Railway Express between various points, and checked on the number of bus, railroad and air passengers between pairs of cities. All of these were appropriate outcroppings for the test of his hypothesis on inverse proportionality, and in all cases the data conform, more or less closely, to his prediction.

Other investigators have used the continuing record of the newspaper for their data. Grusky (1963) wanted to investigate the relationship between administrative succession and subsequent change in group performance. One could manipulate leaders in a small group laboratory, but in addition one can go, as Grusky did, to the newspapers for more "natural" and less reactive intelligence. Sticking to the sports pages, Grusky learned the performance records of various professional football and baseball teams, as well as the timing of changes in coaches and managers. "Does changing a manager make a difference, or is it the meaningless machination of a front office looking for a scapegoat?" This old sports writer's question is a group dynamics problem, phrased through the stating of two plausible rival hypotheses.

This public record characteristic of the newspaper also allows linguistic analysis. If verbal behavior really is expressive, then one should be able to study a President's position on issues by studying the transcripts of his press conferences. Those answers on which a president stumbles in syntax, or prefaces his answer by a string of evasive dependent clauses, may be symptomatic of trouble areas. Similarly, those questions which receive unusually long or short replies may reflect significant content areas.

Analysis of transcripts such as these can be very difficult, and often not enough substantive knowledge is available to rule out alternative hypotheses. A President is briefed on what are to be the likely topics of reporters' questions, and he has an opportunity to rehearse replies. The setting is not a nonreactive one, and the awareness of his visibility and the import of his answers may influence their content and form. One must also make each President his own control. The verbal styles of Eisenhower, Kennedy and Johnson varied so greatly that any verbal index of syntax, glibness or folksiness must be adjusted for the response tendencies of the individual President.

Less august reporting, that of the society news, served for James (1958) as evidence of community structure. The reporting of social events is highly selective, of course, and most useful for studies of the upper class (cf. Coleman, 1964). The court tennis victory of a truckdriver does not get reported, nor the visit of his wife to Dubuque for the weekend.

Comparison across different cities might be differentially affected by shifts in the selectivity of society editors. It is a good assumption that the size of the city in which the paper is printed will be related to the selectivity of its social news: the larger the city, the greater the probability that a smaller segment of the city's population will appear in the society pages.

Middleton (1960), mentioned earlier, conducted a longitudinal study of the fertility values in magazine fiction, linking it to the actuarial fertility figures. This research suggested that the media, if carefully selected, can serve as a mirror of the society's value--or at least of some selective elements within the society.

How are psychologists, psychiatrists and other psychologically oriented personnel differentiated by the society? One can ask people of course, and one should. But of value, too, is a study of what the mass media contain on the question. Ehrle and Johnson (1961) plucked 4,760 cartoons from six different consumer magazines--all of which pictured psychological personnel. Their evidence suggests no substantial differentiation among the groups. This finding could be further tested by observing psychologists at cocktail parties and noting how often they are asked, "Now how is a psychologist different from a psychiatrist?" Or, one could ask the psychologists to relate their cocktail party experience.

Photographs from consumer magazines could be used to test the speculation of Bush and Londen (1960) that "differences in modes of dress within a particular society are indicators of differences in social roles and self-concepts of members of that society." Whether editorial selection will vitiate any valid comparison is something that would have to be tested, but much is contained in photographs that could be applied to social research.

Sebald (1962) conducted a cross-cultural study of German and American national character, employing the lyrics of songs published in songbooks. His content analysis compared the books along the dimensions of death, authority and manhood. Whether the songs are representative of the society, or whether any collection of lyrics will offer a sensitive enough instrument to pick up the nuances of cultural differences in attitudes toward death is open to debate. It may be an insensitive instrument, but if its findings ran clearly contradictory to those from other data sources, one would have to demonstrate the methodological infirmities that produced the difference or question the other findings.

One such infirmity is that the songs may not be known by more than a small segment of the society under study. Or, some widely popular songs don't get published. A psychoanalytically oriented investigator might see in lullaby collections an opportunity for cross-cultural study of maternal hostility (do Korean lullaby collections contain, "Rockabye baby . . . down will come baby, cradle and all?"). But do mothers have a repertoire of lullabies, do they know the songs in the book? If they did, and the lullabies had a gradient of hostility, it would be a useful individual datum to know her response hierarchy.

These have been examples of partial evidence contained in the mass media of attitudes or social structure. Even the most vitriolic critics of television commercials will admit that the media themselves are a force within the society for socialization of the young and attitude change of the old. Thus, they justify study. Steiner (1963), for example, demonstrated the salience of television for a national U.S. sample by showing the extreme alacrity with which sets are repaired. The old saw that a cigarette was the only object which would cause someone to get out of bed, and go into the night to buy has lost its singularity with the coming of television.

A general review of media effects is contained in Klapper's (1960) excellent summary of the literature. Other useful reviews are Hovland (1954), Klapper (1957-1958) and Lang and Lang (1959). Work has been conducted on the effect the media have on each other (Stouffer, 1962; Parker, 1963), but more exists on the political bias of the media. A content analysis of press reports on presidential campaigns is as predictable as the campaigns, but a fine, relatively unused, source is photographs. The element of editorial selectivity which is a contaminant in some other studies becomes the center of analysis. Editors have a large pool of photographs of a candidate from which to pick, and the one they eventually choose is a revealing piece of intelligence. One of the writers has noted this in American political campaigns, and Matthews (1957) has suggested that it is a phenomenon that might be studied across societies. Writing of the British press, he states:

> They (photographs) can be made to lie . . . as Lord Northcliffe was one of the first to discover. When he was using the Daily Mail to try to get Asquith out as Prime Minister and Lloyd George in, he once issued this order: "Get a smiling picture of Lloyd George, and underneath put the caption 'Do It Now,' and get the worst possible picture of Asquith and label it. 'Wait and See.'"

As in all these studies of the running record, there is the opportunity for time-series analysis. One could learn if a medium is changing in its posture to a candidate (or potential candidate), and observe when. This last information, on the time of modification, can help to validate other sources of data on the medium's attitude.

The selective practice has been so prevalent over time that it is likely to have little "instrument decay" to invalidate time-series analysis. It continues now, and offers some possibilities for interesting new research. A television interviewer recently told Malcolm X, the Black Muslim leader, that he was surprised at how much Malcolm smiled. The Negro leader said that no newspaper ever carried a smiling picture of him. For less extreme, but still marginal leaders, what is the pattern across time and regions? Of equal interest would be the photographic strategy toward Richard Nixon after the 1960 presidential campaign. In a caustic and venomous statement immediately after the campaign, Nixon castigated the press for what he perceived as its anti-Nixon stance. To casual observers, there was an immediate shift in the favorability of Nixon photographs carried by the press. An easy hypothesis to verify.

Tannenbaum (1959) studied press bias another way--analyzing the verbs that appeared in sports headlines. He asked how many runs in baseball equal a "romp" or how many points in football equal " X rolls over Y." In addition to providing descriptive information on the empirical limits of such verb usage, he established a home town bias. The same margin of runs might equal "Sox edged 8-7" or "Sox Bludgeon Yankees 8-7."

Winship and Allport (1943) were unconcerned with selling newspapers, but they did want to know something of the effect of positive and negative stimuli. Their study was conducted during the early years of World War II, and they were opportunistic enough to exploit the "victories" and "withdrawals" blazoned on headlines. Do potential readers buy more or less papers with positive headlines? For the measure of effect, they took the street stand sales of newspapers from four major cities--ignoring the total number of copies sold. This measure properly excludes the relatively invariant element of home delivered newspapers.

Only one paper in the four cities showed any significant sales difference traceable to the optomistic and pessimistic headlines, the Chicago Tribune.

Optimism is a central element in another study--Griffith's (1949) original and adroit research on horse race bettors. The newspapers supplied the odds, results, and payoffs for 1,386 horse races run in the spring and summer of 1947. His hypothesis is worth quoting:

> If the psychological odds equaled the a posteriori (odds) given by the reciprocal of the percentage (of) winners, the product of the number of winners and their odds would equal the number of entries at each odd group after correction of the odds had been made for loss due to breakage and take. If the product exceeds the number of entries, the psychological odds were too large; if the product is less than the number of entries, the odds were too small. (p. 292)

The results, which should receive some distribution beyond the archives of the American Journal of Psychology, suggest that long shots and favorites are overbet, while not enough money is put on middle range odds horses.

The DeCharms and Moeller (1962) study of patents and achievement imagery in children's readers was mentioned above. Others have worked with the content of books, attempting to puzzle out the popular mystery of why some books sell and others don't. Berreman (1940) gathered data on: the sales of books as reported by the New York Herald Tribune book review section, publicity outlays for individual titles and a content analysis. He analyzed 60 titles and grouped them into twelve classes by setting, theme and treatment of theme. His conclusions were, in an oversimplification of a complex Ph.D. dissertation, that content was more important than publicity. His findings suggest that content probably depressed the sale of the well promoted poor sellers and "made" best sellers which had only feeble marketing efforts.

Kappel (1948) took book club selections, a random sample of new titles and best sellers for the years 1926 through 1941 and checked reports from the Book Review Digest on their judged literary quality. A favorability index was plotted for each class of books for each year, permitting both a sequential and aggregate analysis.

Such sales data on individual items within the mass media permit the testing of hypotheses on the way in which items of popular culture rise and fall in favor. One hypothesis now being tested by Webb and R. Armstrong is that sales of a book decline logarithmically once the sales peak has been reached. Data collected so far show some preliminary weak support for this hypothesis. Webb (1962c) demonstrated in a similar study that ratings of television westerns declined logarithmically once they started to slide in popularity.

If one wanted to write formula novels, he might well consult Harvey's (1953) research on the content of best sellers. Harvey relied on secondary sources (the trade magazine Publishers Weekly and Alice Hackett's Fifty Years of Best Sellers,1895-1945) for a list of best sellers for the years 1930 to 1946. A control group of novels that didn't make the best seller lists was prepared, and 30 pages of each novel from both groups were content analyzed for sixteen content factors. These were evaluated by the discriminate functions technique to devise multiple combinations of variables for the predictions of sales.

Soviet writings on psychology could hardly be classed as mass media, but the findings of O'Connor's (1961) research are of substantive interest to any social scientist. O'Connor observed the amount of partisan philosophical (as opposed to empirical) content in Russian journal articles and notes "a tendency to move away from philosophical prolegomena in journal articles and towards a direct discussion of experimental material" (p. 14, cited in Brozek, 1964).

Data Transformations and Indices of the Running Record

Of all the different classes of data treated in this monograph none have so great a need for transformation as those cited in this chapter. Because the data are drawn from continuous records which typically extend over long periods of time, all the extraneous events of history are at work to threaten valid research comparisons.

Perhaps the most obvious of these is the change in the size of the population. The increase in the number of people has meant that the absolute values of actuarial and allied data are relatively useless for comparative purposes. In studies employing election records, for instance, the absolute number of votes cast provides an inadequate base for most research purposes. It gives Mr. Nixon little comfort, we are sure, to know that he gathered more votes in 1960 as a loser than did any preceding winning candidate except Eisenhower. Similarly, the absolute number of entries associated with population level have changed over time. This secular trend in the data is often best removed. Thus, Ianni (1957) had to construct a relative index of residential mobility over time, and DeCharms and Moeller (1962) transformed patent production to an index tied to population.

Time also works its effect by a change in the composition of a critical group. The number of congressmen in the House of Representatives may stay relatively stable over a long time period, but the characteristics of these congressmen change--and in changing, produce a set of rival hypotheses for some investigator's pet explanation of a research comparison. A recent Supreme Court ruling on reapportionment of the House (to more adequately reflect population distribution) will mean substantial changes in the aggregate voting behavior of the House, as well as influencing the decisional setting for all congressmen, those there before the change and the new members.

With known changes in composition, it may be necessary to segregate research findings by time periods in which relatively homogeneous external conditions held. This is a grosser correction than the more continuous correction possible for data linked to population. Even with population, though, the only thoroughly reliable data--the census totals--are produced only once every ten years. The accuracy of intervening estimates, whether from the Census Bureau itself or the highly reliable _Sales Management_ magazine, are high but still imperfect.

The frailty of individual sets of records, which will be discussed below, has caused many investigators to employ indices which combine several different types or units of information. The adequacy of such combinations rests, of course, on the degree to which the component elements are adequate outcroppings of the research hypothesis, as well as the degree to which appropriate weights can be assigned to the elements. Setting these questions aside, however, it is apparent that combining indices must be employed when an investigator lacks so precise and subtle a theory to predict a single critical test, or, when the theory's precision is adequate, no data exist for the critical test. For Angell's (1951) purpose in studying cities, there was no acceptable alternative to first transforming welfare expenditures and crime incidence into indices and then relating them by another index. And for MacRae (1954 b) and Riker and

Niemi (1962), the unstable nature of a single vote by a congressman forced them into constructing indices of samples of votes, which were hopefully a less ephemeral source for comparisons. MacRae needed a "liberal index," Riker and Niemi, and "index of coalitions." Because the individual unit was highly suspect as a sampling of the critical behavior under study, the sampling had to be expanded whenever this occurs. There occurs too the attendant questions of how the units are to be stated, weighted and combined.

One of major gains of the running record then, its capability to study a hypothesis as external conditions vary over time, demands that the investigator consider all transformations possible precedent to making comparisons, and also decide whether indices will yield a more stable and valid base for his hypothesis testing. This requirement is not as pronounced in the discontinuous archival records cited in the chapter that follows nor among the observational and physical evidence methods.

Over-All Evaluation of Running Records

It should be obvious that we prize the potential for historical analysis contained in running records.

> The best fact is one that is set in a context, that is known in relation to other facts, that is perceived in part in the context of its past, that comes into understanding as an event which acquires significance because it belongs in a continuous dynamic sequence. . . (Boring, 1963, p. 5).

If a research hypothesis, particularly for social behavior, can survive the assaults of changing times and conditions, its plausibility is far greater than in a method which strips away alien threats and evaluates the hypothesis in an assumptive, one time, test. Validity can be inferred from a hypothesis' robustness. If the events of time are vacillating, as they usually are, then only the valid hypothesis has the intellectual robustness to be sustained while rival hypotheses expire.

One pays a price in such time-series analysis, the necessary price of uncertainty. We again agree with that gentle stylist Boring (1963):

> The seats on the train of progress all face backwards; you can see the past but only guess about the future" (p. 5).

A hypothesis might not hold for anything but the past, but if the present is tested, and a new, possibly better, hypothesis produced, those same running records are available, as economical as ever, for restudy and new testing.

For all the gains, however, the gnawing reality remains that archives have been produced for someone else and by someone else. There must be a careful evaluation of the way in which the records were produced, for the risk is high that one is getting a cut-rate version of another's errors. At the beginning of this chapter, we detailed the operating questions of selective deposit and selective survival of archives. Both these contaminants can add significant restrictions to the content and contributing populations of the archival materials. In the discussion of individual research studies, we have noted how roll call votes, marriage records, reports of congressional speeches, letters to the editor, crime reports, and other records are all subject to substantial population and/or content restrictions in their initial recording. To a lesser degree, the selective survival of records can be a serious contaminant, but in certain areas, such as politics, it is always a prime question.

Those contaminants which threaten the temporal and cross-sectional stability of the data are controllable through data transformation and indexing methods--if they can be known. Happily, one of the more engaging attributes of many of these records is that they contain a body of auxiliary data which allows the investigator good access to knowledge of the population restrictions. We have noted this for the absentee contaminant in congressional voting and the selective choice of cases in judicial proceedings. With the actuarial material on birth, marriage and death, it is often possible to find within the records, or in associated data series such as the census, information which will provide checks on the extent to which the research population is representative of the universe to which the findings are to be generalized.

If the restrictions can be known, then it is possible to consider the alternative of randomly sampling from the body of records, with a stratification control based on the knowledge of the population restriction. This is feasible for many of the records we have mentioned because of their massiveness. Indeed, even if no substantial population contaminants exist, it is often advisable to sample the data because of their unwieldy bulk. Since usually they can be divided into convenient sampling units, and also frequently classified in a form appropriate for stratification, the ability to sample archival materials, particularly those in a continuous series, is a decided advantage for this class of data. The sampling of observations, or of traces of physical evidence, is markedly more difficult.

The population restrictions are potentially controllable through auxiliary intelligence; the content restrictions are more awkward. For all the variegated records available, there may still be no single set, or combination of sets, that provide an appropriate test of a hypothesis. This is particularly true for process phenomena in such areas of study as learning. One cannot go to a city hall and find a Department of Retroactive Inhibition--the way one can find a Department of Vital Statistics. The student of learning may find some pertinent data in Vital Statistics, but it will be less visibly applicable to his research interests than it will be to the student of fertility rates or suicide.

Something of this content inability is reflected in Walter Lippmann's (1955) discussion of the "decline of the west." Lippmann writes of the turn of the century when

> The public interest could be equated with that which was revealed in election returns, in sales reports, balance sheets, circulation figures, and statistics of expansion. As long as peace could be taken for granted, the public good could be thought of as being immanent in the aggregate of private transactions" (p. 16).

Yet many of the studies reported in this chapter have revealed the power of insightful minds to see appropriate data where associates only see "someone else's" records. There is little explicit in patent records, city water pressure archives or children's readers to suggest their research utility. It required imagination to perceive the application, and a willingness to follow an unconventional line of data collection. Imagination cannot, of course, provide data if none is there. Our thesis is solely that the content limitations of archival records are not so great as the social scientist bound by orthodoxy thinks.

There is no easy answer on the degree to which reactive measurement errors exist in running archival records. These are second-hand measures, and many of them are contaminated by reactive biases while others are not. The politician voting on a bill is well aware that his action will be noted by others; he may not be aware that an observer in the gallery jotted down the tic in his left eye when his name was called to vote. The records contributed by the person or group studied--the votes, the speeches, the entries written for directories--are produced with an awareness that they may be interpreted as expressive behavior. Thus, those errors that come from awareness of being tested, role elicitation, response sets and from the act of measurement as a change agent are all potentially working to confound comparisons. With other data, such as the reports of presidential press conferences and census figures, the investigator has the additional bias of possible interviewer error passed along.

For data collected by a second party, by someone other than the producer (birth and death records, weather reports, power failures, patents and the like), the risk of awareness, role or interviewer contaminants is low. The problem here becomes one of instrument decay. Has the record keeping process been constant or knowably variant over the period of study? It may be that response sets on the part of the recordkeepers, or a change in administrative practice, threatens valid comparisons across time periods or geographic areas. To know of this variation is extremely difficult, and it represents one of the major drawbacks to archival records.

In summary, the running archival records offer a large mass of pertinent data for many substantive areas of research. They are cheap to obtain, easy to sample, and the population restrictions associated with them are often knowable and controllable through data transformations and the constructions of indices. But all content is not amenable to study by archival records, and there is an everpresent risk that reactive or other elements in the data producing process will produce selective deposit or survival of the material. Against this must be balanced the opportunity for longitudinal studies over time studies in which one may test a hypothesis by subjecting it to the rigor of evaluation in multiple settings and multiple times.

Chapter 4

Archives II: The Episodic and Private Record

In the preceding chapter, we outlined the joys and sorrows of those archives on which there is a running time record. Here we will continue our discussion of archives, but center on those which are more discontinuous and typically not a part of the public record. Such data are more difficult to come upon than the public records, unless the investigator is affiliated with some organization producing the material. The insurance sales of a casualty company, the nurse's record on a bedside clipboard and last year's suicide notes from Los Angeles are more available to the "inside" investigator than they are to the curious outsider. But if these records are more difficult and costly to acquire than public records, they often provide a gain in content specificity. The amount of irrelevant dross commonly declines as an investigation is limited to a particular set of privately produced data.

We have already mentioned the risks to validity inherently present in archival records. The main analytic difference between the records mentioned in this and the earlier chapter is the common inability to make longitudinal analyses of the private data. Sometimes security is the reason, sometimes the data are stored for shorter periods, sometimes financial and labor costs preclude an analysis over time. Whatever the cause, this is a major loss. The best defense against it is to find a related set and combine both--one continuous and the other discontinuous--for a more textured series of comparisons.

Some of the data in this chapter is episodic in character, but complete in reporting; many sources do maintain long and accurate record keeping systems. The military is one such source, and Lodge (1963) has conducted a provocative correlational study with U. S. Navy records. All those who have learned his results on air crashes tend now, as passengers, to squint studiously at their pilots' height. Dipping back into the Navy records, Lodge collected reports of 680 jet plane accidents, and then searched other records for the height of the pilots. He learned that men exceeding the average height of 72 inches had significantly more accidents than their shorter contemporaries. This may be traced to the construction of aircraft cockpits--which are designed for those of average height.

We have divided the sources into three gross classes: sales records, institutional records and personal documents. All three are potential substitutes for direct observation of behavior. This is most obvious with personal documents, where the unavailability of a source may force the investigator to use whatever alternatives are available. But the sales and institutional records may work in the same way, and can broaden the scope of an investigation which is primarily based around observation. They may fill in holes present in an observational series, or used to produce a broader sampling of the behavior under study.

Chadwick Alger (1962) once suggested to us that it would be profitable for a political scientist to sit in the Delegate's Lounge of the United Nations and observe how much whisky was downed. By keying the consumption rate to action before the UN, Alger felt that an index of tension might be developed. In such a setting, the sales records of the bar might be an even better measure, for they are less amenable to instrument decay errors, and permit a closer noting of type of drink ordered (Scotch, Canadian, Cuba libres, etc.).

Brown (1960) suggested that soap consumption records be substituted for ratings of the cleanliness of institutionalized patients--ratings which are, after all, observation one step removed. Brown points out two ways in which soap usage could be measured: a liquid soap could be measured by reading the level of liquid in the dispenser each day or, with bar soap, by a measure of the amount of water displacement at the beginning and end of the period studied.

Two other examples, both employing records of whisky consumption, will illustrate the substitution of sales records for observation. Hotel and restaurant records could be employed in a comparative study of occupations. One can observe members of an occupation and attribute traits to them, but valuable auxiliary information might come from the records on drink consumption and petty thievery in convention hotels. Do anthropologists take more soap and towels away with them than do mechanical engineers? Such an analysis is posited on the assumption that those who attend conventions (and who stay, steal and drink in convention hotels) are a representative sample of the profession.

Hillebrandt (1962) sought data on the sale of alcoholic drinks at Chicago airports in his study of passenger anxiety produced by aircrashes. He failed to use the proposed data because of the insensitivity of the instrument. At that time, the major Chicago airport had been recently completed, and the construction of bars had not caught up with demand. Thus, there was negligible variance from day to day in sales, and the small amount that existed seemed to be based upon the bartenders' speed, and not on exterior factors.

Sales Records

In a society so oriented to marketing and record keeping as ours, sales data abound for study of a varied body of content. As noted above, Hillebrandt (1962) did not get to use bar sales as a measure. He continued on with his study, however, and used the volume of air traffic (air ticket sales) as his measure. With a complete set of data over time, he was able to transform the data, correcting for systematic sources of variance irrelevant to his hypothesis. He partialled out, for example, the seasonal variation in air traffic, and accounted for the secular changes in traffic level at Chicago's two major airports. The residual material demonstrated that crashes had only a very short term effect on traffic. Just as in the bar sales, there is some rigidity in his data, rigidity which tends to blunt comparisons. A certain number of people have no alternative but to fly, regardless of how dissettling was a major crash the day before. Campbell and Webb are continuing this analysis with a complementary measure. With the same exogenous variable, they are plotting the number and dollar value of trip insurance policies taken out by travellers at Chicago airports. Buying trip insurance is a low cost and simple behavior which should index anxiety of air travellers. With the bar facilities now expanded to exceed demand, it will be possible to perform an analysis containing three sales components: bar sales, trip insurance sales, and ticket sales. Each of these must be corrected for the systematic biases present, but they should provide a more sensitive and nonreactive set of possible outcroppings of anxiety than any single variable study, particularly one based on the interviewing of travellers.

Insurance, a paid for hedge against risk, is an admirable measure of the effect of disaster. Just as one can examine trip insurance sales and link them to crashes, one can examine the timing of casualty insurance sales and link them to the occurrence of hurricanes or tornadoes. Add to this sales of life insurance (compared to the time of death of close friends or relatives) and one has a three-way index of the general effect of disaster

These same data could be used to further test Zipf's (1946) hypothesis. Is the amount of casualty insurance taken out inversely proportional to the distance from the disaster? How the mapping of insurance underwriting compares to a meteorological map of tornado probability would be a necessary control. It might be that the hypothesis holds only in areas which have had tornado experience. This would give support to Zipf's hypothesis, but even greater support would come if a significant amount of insurance were written in proximate areas with little or no tornado experience. What, in brief, is the nature of the generalization of effect?

How unlikely a source of research material is the sale of peanuts! Yet, continuing in the vein of study of anxiety or tension, peanut sales are a possibility that should be systematically explored. An anecdotal report appeared in the Chicago Sun-Times from the concessionaire in that city's baseball parks. He casually observed that peanut sales after the seventh inning of a game are significantly higher than earlier--but only during a tight game. If the game is one-sided, there is no late inning increase in peanut purchasing. Is this a sound, nonreactive measure of involvement or tension? It may be, but it illustrates the problems associated with such archival measures--one must pay special heed to rival hypotheses. One hypothesis is that fans, during the increasingly tense moments of the late innings, absentmindedly lean over and compulsively crunch their way through more peanuts than earlier.

But one should look at population restrictions. It may be that the finding (if it may be legitimatized as such) on the increase coming only in tight games is an artifact of selective attendance, and not tension. There is a hyperbolic curve of attendance in a one-sided game. A substantial number of fans usually arrive late, and another substantial group leave early if the game appears to be already decided. The population potential for peanut purchasing is thereby variable across innings; the effect of the tight game is to reduce the early departures and provide a larger base for late inning sales. This is only a partial explanation, to be sure, for most of the fans will have arrived by the third inning. If one wanted a finer test, a simple correction would be to transform the peanut sales into unit sales per X thousand fans per inning--a transformation possible by clocking turnstile movement in and out.

Another test of the hypothesis, using the same data, is possible. Unfortunately, the raw data has not been made available and we can report no findings. The tension hypothesis would get strong support if in a population of one-sided games, fans left in substantial numbers early and sales stayed stable from the middle through the later innings. This would be reflected in an ascending consumption curve if correction were made for those in attendance.

Not to beat peanuts to death, but one other element of population restrictions remains to be considered. We have examined only the issue of the absolute number of fans available in the park to buy the peanuts. Is there any plausibility to the notion that those who leave early are more or less devoted to peanuts? One might determine this by interview at the exit gate, or by looking for traces of peanut shells in vacated seats.

Sales data can also be used to infer popularity and preference. The impact of Glenn's orbital flight was evidenced by record breaking sales of the commemorative stamp issued to mark it (New York Times, 1962). Similarly, the sale of commemorative Kennedy stamps and the great demand for Kennedy half-dollars after their issuance, as well as all the special books and the renaissance of Profiles in Courage on the best seller lists provide persuasive evidence, if any were needed, of the man's influence on public thinking.

Stamps, their sales and pictorial content, are a useful but unused bit of expressive intelligence. An analysis of the illustrations on stamps may give indications of the state of political opinion in the nation. What does an analysis of illustrations printed during the early years of the fascist regimes in Germany and Italy show? It has been suggested that stamp illustrations presage aggressive political action. Perhaps somewhere a philatelistic psychologist has prepared this study, but we have been unable to come across it.

A possible flaw in such analysis is a potential datum for another topic of study. Each nation does not print its own stamps. Many former colonial territories continue, as a cost saving device, to use second-hand engravings and printing facilities from the former governing power. The illustration may thereby reflect an economic and not a political decision. But whether or not a former colonial territory still relies on the old power is itself a clue to relations between the two. One could, for example, compare former British and French colonies in Africa. Or, compare new nations within what was a single colonial area. Guinea prints its own stamps; Mali buys hers from France.

Asius (1962) used the sales of film to index the relative popularity of different camera types. The sales of the camera types proper would be one measure of popularity, but the film sale measure provides a clue to the preference of those who own more than one type.

The diffusion of information among physicians was the topic of Coleman, Katz and Menzel's (1957) research. Instead of the more standard, and reactive, tactic of interviewing doctors, they elected to go to the pharmacy records to note which doctors prescribed what drugs when. Sampling at intervals over a 15-month interval, they related the physician's adoption of new drugs to his social network. Such hardnosed data can provide a useful check on interviewing data, providing that the effect of collecting such records does not alter the behavior of the record keepers. This is a very implausible risk with drug prescriptions, but a reasonable one when one deals with less legally controlled records. The danger is not so much in the masking of information as it is in better information. The record keeper may perform a more conscientious job because he knows that his work is being put to some use. If the investigator using such records stresses the greater glory to man that will come from the record keeper offering his cooperation, he may actually be increasing the risk that the instrumentation process will change--thereby threatening the validity of comparisons over time.

The social network variable employed by Coleman, Katz and Menzel can be measured by other methods than the standard interview or sociogram. To study the extent to which interaction among different departments of a university took place, one could go to the orthodox procedures. In addition so humble a document as a desk calendar might be checked. This record can provide information on who lunched with whom, with what degree of frequency and across what departments. Not everyone notes such engagements (a population restriction) and the desk calendar is not a likely source for learning of other engagements, such as social dinners or meetings so regular that they don't have to be noted (content restrictions). Staying just with the lunch dates, one route to learn the character of the restrictions would be to enlist the aid of waiters in faculty clubs. The reader can conjure up further objections to this assistance.

Drug sale records are in common use by ethical pharmaceutical houses to evaluate the sales effort of their detail men. Both the houses and the detail men have learned that the verbal statement or the observed enthusiasm of a doctor for a new drug is a highly unstable predictor of what he will prescribe. If records are available for checks on self-reports, they should by all means be used. Such checks are particularly useful when the data are being continuously produced by the same subjects--for then a correction can be applied to the self-reports. If the assumption can be made that the character of the error is constant over time, it may not be necessary to run both sets of data concurrently.[1]

---

[1] Something of an analogous correction can be seen with economic forecasters. Firms which employ the forecasters commonly compute the response set characteristic of the forecaster (optimism or pessimism) and apply a secret correction to his periodic predictions. This is a rather interesting data transformation, since its existence shows the set rigidity of the forecasters. They, too, know whether they have been over- or under-estimating, yet that knowledge does not produce potent enough feedback to overcome the response set.

Boring (1961) has detailed how a similar type of response error was an early *cause celebre* in astronomy. Differences in reaction times among astronomical observers became known, and the phrase "personal equation" was coined to describe the bias. In an evolving history, the contaminant existed but was unknown, became known and the cause of study for the purpose of eliminating its biasing effect, and then became the substantive material of a large body of psychological research.

The sale of stocks was used by Ashley (1962) in a study interesting for what it says of positive and negative reward. Isolating firms which announced an unexpected dividend or earning statement, either up or down, Ashley traced the stock prices following the announcement. An unexpectedly high dividend influenced the price of the stock for about 15 days, while an unexpectedly low dividend or earnings statement had an effect for only about four days. There are other places to measure extinction than in a Skinner box.[1]

A more common use of sales records (but still surprisingly uncommon) is as a measure of propaganda effectiveness. Within advertising particularly, there has been extensive writing on the inadequacy of survey methodology to predict the advertiser's major criterion of sales. Henry (1958) gives numerous examples of the discrepancies between respondents' reports and sales figures, indicating that the reasons a consumer gives for buying a product cannot be relied upon. Reactive measure thus suspect, other methods of consumer preference measurement must be found.

Henry mentions an example of a controlled sales experiment on "shelf appeal." One brand of candy is sold in three types of wrappers, with variables such as shelf position and number of units displayed controlled. The dependent variable is the sales level of the different packages.

There is an ideal experimental medium for advertising research in direct mail sales efforts. Lucas and Britt (1963) comment upon a number of studies which have varied such elements as the color of paper, inclusion of various incentives and stamping a return envelope by hand or by meter. They also cite the example of a department store which might send out a monthly statement to customers containing one of ten variants of an advertisement selling a specific product. The average return per layout would then be determined by subsequent sales.

---

[1] Cf. Winship and Allport's (1943) study on newspaper headlines and Griffith's (1949) research on horse race odds--both studies of optimism-pessimism. Hamilton (1942) has also conducted an interesting content analysis on the rise of pessimism in widely circulated Protestant sermons.

The large mail order houses (Sears, Wards, Spiegel's, etc.) regularly conduct research on different thematic appeals. This is easily performed by varying the content of the appeal and simply counting the returns attributable to the different sources. In a revealing finding, one of these houses found that advertising a self-riding lawnmower (which costs several hundred dollars) as something of an adult's toy dramatically outsold an appeal which argued its superior functional merits.

Over 30 years ago, Zeisel (1932) studied the impact of the depression by noting the level of grocery sales. The same measure is used now to evaluate the efficacy of different sales themes. A before-after design was used by the National Advertising Company (1963) to learn the effect of large outdoor signs placed in the parking lots of three shopping centers. They used store audits to determine the sales of products advertised on the signs and then compared these data to sales from a control sample of equivalent stores. Another method of "pre-testing" advertising themes is to post treatments of the theme in prominent places within a supermarket and then observe sales. This has been done by comparing sales attributable to a single theme against a control of past sales, and also by employing multiple themes and comparing one against the other. In this latter approach, close attention has been paid to time sampling problems, so as to protect the equivalence of the populations exposed to each version.

One could similarly experiment with vending machines, although we do not know of any such research. By random assignment of the display of experimental cigarette packages to machines or by systematically varying exhortatory messages over machines, one could employ level pulling as the effectiveness measure.

Aside from the commercial applications of such research, direct mail advertising, vending machines and the like offer a fine natural laboratory or the study of persuasion. Mindak et al. (1963) used the sales of tickets to parades and to a civic aquatennial as one of their several measures of the effect of a newspaper strike. Roens (1961) reported on the use of different combinations of media to carry the same propaganda theme (for Scott paper), and in Berreman's (1940) study of factors affecting the sale of novels, he demonstrated that publicity elements had little effect.

DeFleur and Petranoff (1959) investigated subliminal persuasion by sales measures. Since the alleged effect was first reported (with no data) as the result of a filmed message, they flashed "Buy Product X" on a screen subliminally for several experimental weeks. The effect of this was measured by the deviation from normal of a food wholesaler's orders. No significant effect was observed. The transfer of such findings to a more general treatment of persuasive phenomena is obvious.

The stuff of commercial persuasion, advertisements proper, have been employed in a number of ways for other purposes. The want ad columns of newspapers have served as economic predictors (Fowler, 1962). Contrariwise, they may provide the data for historical analysis. The Security First National Bank of Los Angeles compiles a regional index of want ad frequency, and reports variance over time in this index coincidental with basic economic data. A management consulting firm prepares an index based only on advertisements for engineers and scientists. Although incomplete data are reported by

Fowler, she reports the firm "likes to consider its index as a leading indicator of how the economy fares, rather than a coincident indicator" (p. 12). That optimistic assessment illustrates the necessity for considering the time linked nature of a measure. The statement was true for a long period of time--from the late 40's possibly up to 1964. But since the employment of engineers and scientists is intimately tied to defense contracts, a major chance in Defense Department policy will throw off employment levels, thus the ads, thus the indicator. Just such a change occurred during Secretary McNamara's administration, and the 1964 advertisement levels should be negatively correlated with the state of the economy.

Assuming that ads reflect values, Dornbusch and Hickman (1959) sampled 816 issues of the Ladies Home Journal and content analyzed advertising to estimate the degree of "other-directedness" displayed. These data tested Riesman's hypotheses on the history of "other-directedness". Two classes of indices were employed: (1) endorsements by persons or groups and (2) claims that a product is related to satisfactions in interpersonal relations.

Singh and Huang (1962) made a cross-cultural study of advertising. They compared American and Indian advertising for similarity and related the findings to socioeconomic and cultural factors. Since they used print advertising, it is important to take account of differential literacy rates in such comparisons. If the advertisers are addressing their messages to literate "prospects" and reflecting the values of those "prospects," the differential literacy rate may interact with magazine readership and prospect status to yield differences between societies that are spurious. This is less a concern with broadcast advertising, although there may be a population in that those who either own or are more available to radio or television receivers vary systematically across countries.

Industrial and Institutional Records

Among the finest work to be found on discussion of multiple methods and the criterion problem has been produced by the industrial psychologists. One rarely finds such attention to the relative merits of ratings v. observation v. performance v. interviewing v. questionnaires v. tests. Guion (1961) has recently offerred an excellent short treatment of this subject, as have Whisler and Harper (1962) and Ghiselli and Brown (1955). But these have not supplanted the singular statements on "criteria of criteria" by R. L. Thorndike (1949).[1]

The number of private records marshalled by the industrial psychologists has been impressive. Amount and quality of output are probably the most frequently used behavioral measures, and are usually expressed in some transformed score--the number of units produced by a worker or department per unit time, the amount of sales per unit time, or the profitability of activity by dollars invested by the firm. The subjectivity of ratings by supervisors or foreman increasingly has moved many of the specialists in this area to pure behavioral measures.

J. S. Adams (1964) has demonstrated what fine research can be undertaken with simple productivity data in his important work on wages. He conducted three experiments "to test how people behave when they are working on a relatively highly paid job for which they feel underqualified (that is, when they feel their pay, or outcome, exceeds their qualifications or input)" (p. 10).

The subjects were students who were hired for part-time temporary interviewing, not knowing they were part of an experiment. In the first experiment, they were paid $3.50 an hour and divided into two groups. The experimental group was led to believe that they were not qualified for the job as interviewers; they "were treated quite harshly." The control group was told they had met all qualifications for the job. With productivity the dependent variable, the experimental subjects produced significantly more than the control subjects.

---

[1] For other articles of methodological interest on criteria see Brogden and Taylor (1950), Gordon (1950), Fiske (1951), Bass (1952), Severin (1952), Rush (1953), MacKinney (1960), and Turner (1960).

In the second experiment, interest was focussed on the relationship between method of pay and dissonance. Thirty-six students were hired for the same interviewing job and randomly assigned to the four conditions:

1. Experimental dissonance condition--Ss paid $3.50 an hour and made to feel overpaid.
2. Control condition--Ss paid $3.50 an hour and made to feel it was an equitable wage.
3. Experimental dissonance condition--Ss paid 30 cents an interview and made to feel overpaid on a piece rate.
4. Control condition--Ss paid 30 cents an interview and made to feel payment was fair.

Adams' hypothesis was supported: "hourly workers in the dissonance condition had a higher mean productivity than their controls, whereas pieceworkers in the dissonance condition had a lower mean productivity than their controls" (p. 13).

In the final experiment, Adams showed that pieceworkers who perceive that they are inequitably overpaid will perform better quality work and have lower productivity than pieceworkers paid the same rate who believe that the wage is fair. The experimental subjects increased their "inputs" on each unit of work, thereby increasing the quality, but decreasing the quantity.

We have detailed this study because of the example that Adams provides of the shrewd hypothesis testing potential for some of the rudimentary, but natural, measures available from institutional records.

These records have their difficulties, and Whisler and Harper (1962) are helpful in illustrating problems in making comparisons across departments or workers. The factors involved in any such comparison of records are numerous and vary from one situation to another--the type of work, physical working conditions, work group cohesion and the like. Ghiselli and Brown (1955) struggle with the problem of appropriate controls for work behavior and they offer a series of possible work standards: group average production, rates of selected individuals known to be "good" workers (with some discounting then applied), experimentally determined times for tasks, and "rational analysis" by people familiar with the problem.

We will review a body of studies with data drawn from the institutional records of companies, schools, hospitals and the military. A good share of these come from industry, but the overlap with other institutions is marked.

Most of the current writers argue for multidimensional criteria. Ghiselli and Brown (1955) give the humble example of a streetcar motorman, indicating a series of proficiency measures:

1. number of collisions with pedestrians
2. number of traffic violations
3. number of commendations from public
4. number of complaints from public
5. number of times company rules broken
6. number of sleepovers (tardiness)
7. number of times schedules are broken
8. number of reprimands from inspectors
9. ratings by inspectors
10. errors reported by dispatchers

As always with multiple measures such as these, the question comes up of the advisability of combining the various measures into a single composite score. Consider in the list above the problem of weighting each of the ten variables. Ghiselli and Brown suggest multiple cutoffs--establishing a minimum performance level for each component element of the index. The way in which the variables are combined, and the final score reached, can be disastrously misleading if some minimum standard is not met on each of the tasks of the job. They offer the highly reasonable example that an airline pilot should be able to land a plane as well as take one off--perhaps not so gracefully, but still with a minimum degree of proficiency.

Another multimeasure study, this one of the ship effectiveness in the Navy, was conducted by Campbell (1956). Rather than rely solely on ratings by the captain or crew members, Campbell examined the ship records for reports of ship inspections, torpedo firings, re-enlistment rates of those aboard, requests for transfer and disciplinary actions.

Re-enlistment rates are a measure of job turnover, and Evan (1963) explored this topic in some student based research. Personnel records were examined to find a possible relationship between job turnover and departmental placement. He reasoned that a high level of interaction on the job with other student-workers would have a stress reducing effect and result in a lower rate of job turnover. He supported this hypothesis by showing that the number of students with whom a worker could interact was a variable tied to quitting.

Knox (1961) added absenteeism to job turnover, and correlated both these records with factors of age, seniority and the distance of the worker's home from the factory. Melbin (1961) also used absenteeism and job turnover in research on psychiatric aids. He compared these data to archival reports on work assignments. The correlational methods limit his ability to establish cause and effect relationships, but he was able to trace a double-directioned effect between changes in work assignments and absences.

Job turnover is an ambiguous measure by its source--sometimes an administrative action by the firm and sometimes an action by the individual employee. R. L. Thorndike (1949) holds that administrative actions should be considered as a discrete class of criteria. Along with many others, Guilford (1956) used the record of pay increases as his measure of the firm's appraisal of the individual. Weitz (1958) suggests the imperfectly correlated measure of promotion within the company, while Jay and Copes (1957) speak of job survival as a criterion. Merely to stay on a job, without being fired, is indicative of an administrative decision that the employee is not that bad.[1]

One could also observe other management actions which reflect the esteem with which an employee is held. The rug on the floor, the drapes on the window, the white telephone, the second secretary and the corner office are all salient cues to an employee's success with the firm.

Other industrial research with absentee records has included Amthauer's (1963) work on social psychological settings and Bernberg's (1952) study of departmental morale. Amthauer took absenteeism as his dependent variable and considered a string of independent variables: motivation, intelligence, persistence and stability, as well as to the individual's relationship with apprentices and with instructors. Note that although absenteeism is not implicitly reactive, the other measures of individual psychological characteristics were. So too for Bernberg, who measured departmental morale by measures with high reactive risk, but matched these against an elaborated set of absentee measures. He took four variants of absenteeism as his variables: unexcused absence of one day or longer, tardiness, absence of less than one day, and trips to the medical unit not resulting from accident or disease. For a review of studies focussed on absenteeism in the industrial setting, see Brayfield and Crockett (1955).

---

[1] Whisler and Harper (1962) also speak of seniority as a criterion and discuss the implications of this in the union-management struggle over definition of criteria. They state that seniority has appeal because of its qualities of objectivity and precision, and that it is not a simple case of the union wanting seniority and the management fighting it. Promotion from within the management group and the high value placed on experience is evidence of the use of seniority within the management tier.

The unions keep books too, and Stuart (1963) was canny enough to use grievance records in a study of racial conflict. He collected 364 verbatim records of the Grievance Board of a large union in the textile industry--data extending over a seven year period. The complaints were analyzed to determine feelings, attitudes and actions against Negro and Spanish-speaking workers who comprised a majority of the union. Their complaints are important evidence, but there is the risk of the bias one notes in studies of political speeches. Although the events occurred in the past, and the investigator did not intrude in the production of the data, the subjects were very much aware that their remarks were "on the record." This perforce limits the degree of generalization possible.

McGrath (1962) also used indirect data in his study of friendship and group behavior. The dependent variable was the competitive performance of rifle teams, plotted against whether the team had given favorable or unfavorable ratings to former teammates and how the team had been rated. Steiner (1964) comments on a possible population restriction:

> "Additional research is needed to determine whether the individualistic orientation which was a boon to rifle teams would also promote productivity in situations calling for cooperative group action" (p. 434).

Stouffer (1949) was among those reporting on sick call data--drawing statistics from the Office of the Air Surgeon on members of heavy bomber crews in the European theatre during June, 1944. The sick call rate was noted and rated to self evaluations of physical condition. Stouffer suggests that the contrast between attitudes toward one's physical condition and behavior with respect to physical symptoms reflected the men's tendency to save up complaints until their tour of 30 missions was completed. He speculates that might have been motivated by a fear of not completing a tour of combat duty and thus running the risk of postponing a return to the United States.

More military records were used by Fiedler, Dodge, Jones and Hutchins (1958) and Fiedler (1962). These studies examined "adjustment" in nonclinical populations--Army units and student populations. For the Army personnel, they obtained sick call data, disciplinary offense ratings and court martial records; for the students, course grades, number of visits to the student health center and student counselling bureau visits. Little relationship was found among these measures of "adjustment," underlying the problem of combining pieces of evidence which look to be similar in reflecting some characteristic.

In a more circumscribed investigation, Mechanic and Volkart (1961) probed sick call visits further, with a population of college students. They showed the frequency of visits to be related to the subject's degree of stress and his tendency to play the sick role. Medical visits are not a pure measure of a single trait, and few single archival measures are ever pure. It is always indicated to employ multiple measures of the same hypothesized "trait" (Campbell & Fiske, 1959). We stress this, because many of the studies which have been venturesome enough to employ nonreactive data have done so to define or classify subjects--often than administering questionnaires and interviews to the subjects. This is a highly appropriate device to combine research methods, but its validity is posited on the accuracy of the trait definition contributed by the initial record. A multiple method approach is the best hedge against error (cf. R. L. Thorndike, 1949).

This approach may be illustrated by the titantic work of Thorndike (1939), although the content is more properly subsumed under government records. In his massive study, Thorndike studied cities by rating 310 U.S. communities on a "goodness scale"--composed of 297 characteristics! Examples of the measures are the average salary of high school teachers, percentage of the city budget allotted to recreational facilities and the infant death rate. As a supplement to the study, Thorndike gave his reading public an opportunity to be nonreactive researchers by presenting a ten-item scale for each reader to use in rating his own city.

The less readily available material of library withdrawals were used by Vernon and Brown (1963), who were interested in measuring the "dynamic information seeking process" among tuberculosis patients. They predicted that the patients of uninformative doctors would differ significantly from patients of informative doctors in the degree to which they sought out disease-relevant information. They indexed this by the proportion of patients in the two groups who took "tuberculosis" and "non-tuberculosis" books from a hospital library. In this study, the authors depended upon the reports of patients to define informative and uninformative doctors. Rashkis and Wallace (1959) demonstrated that the researcher does not have to depend upon self-reports on degree of attention paid by medical personnel. They observed the notes made by attending nurses on a patient's bedside record, notations that were both informal and required. The attention paid to the patient was measured by the frequency of these notes per patient. This helps to circumvent the possibility that perception of how much attention is being paid to one by medical personnel is heavily contaminated by the patient's degree of illness. The same amount of attention would probably be differentially perceived under pre-operative, post-operative and about to be released states. The bedside chart may be more trustworthy.

Schwartz and Stanton (1950) suggest a study of the social situation of the hospital combining observational and archival materials. The observation is of negative incidents in a ward and the archive is a measure of incontinent behavior: the amount and type of laundry done for that ward. In their exploratory observational study, they kept complete records of the patients' daily activities, and were able to determine [illegible] connections between certain types of negative incidents and incontinent behavior. They further suggest that the laundry record could be useful in establishing the effects of changes in therapeutic methods within a ward.

Parker (1963) worked on the effect of the introduction of television by studying library records in a way very similar to that of Vernon and Brown. His topic was the differential effect of television on the reading of books. A time series was prepared of withdrawals, by type of book, from libraries in a series of Illinois cities--both before and after television came to town. In one of these findings that look like common sense (after the research is read), he learned that the withdrawals of nonfiction titles was unaffected, but that there was a significant drop in the withdrawal of fiction titles.[1]

---

[1] Jones (1960) used the extent of library facilities and personnel as the "index of progressivism" of a group of 154 Illinois towns. He corrected the raw figures by applying census data on community size, social class, race, occupation-income, population age and rate of community growth.

Both Parker's and Vernon and Brown's studies have a possible confounding in subject selection. The tuberculin patients may easily have varied on some other characteristic than the informativeness of their doctor. Putting aside the patient designation of informativeness, doctors are not randomly assigned to patients, and the differential populations of patients by ward may have accounted for some of the differences in reading habits. The cities employed by Parker received television after the "freeze" on new station construction was lifted by the Federal Communications Commission. Thus, they may have varied significantly along other dimensions from those cities which received television early. Just how this contaminant would predict a differential withdrawal of fiction and nonfiction titles, however, is difficult to see. Reasons can be put forward, but they lack plausibility. It is the point stressed earlier: content restrictions are not damaging to internal validity unless they operate differentially on the data being compared.

Written Documents

The last major class of more or less private archives is that of personal documents. These have been more the bailiwick of the historiographer than the behavioral scientist, but a number of notable studies have been performed using personal documents. Cox (1926), in Terman's Volume II of Genetic Studies of Genius, used documentary evidence of all kinds, and, on the history of science side, we have Terman's early (1917) study estimating Galton's IQ. Centering on records of Galton's prowess between the ages of 3 and 8 (he could read any English language book by five and knew the Iliad and Odyssey by six), Terman compared them with the ages at which other children are able to accomplish the same or similar achievements and estimated Galton's IQ to be not far from 200.

There have been important methodological works, such as G. W. Allport's (1942) monograph, *The use of personal documents in psychological science*, and facilitating method papers, such as Dollard and Mowrer's (1947) system to determine the amount of tension in written documents by a "Discomfort Relief Quotient."[1]

But for all this, written documents have been another of the underdeveloped data resources of social science. In the examples that follow, we will cite some of the major studies using written documents and illustrate some of the rival hypotheses coincident with them.

One could not think of letters as a research source without bringing to mind Thomas and Znaniecki's (1918) classic study of the Polish peasant. Letters sent between Poland and the United States were one of the major elements in a data pool that included autobiographies, newspaper accounts, court proceedings and the records of social agencies. Rather by happenstance, Thomas learned that there was an extensive correspondence between the two countries and that many of the letters were being thrown away. From this lead, advertisements were placed which offered to pay for each letter produced.

There are, to be sure, substantial questions about the population and content restrictions in the letters Thomas and Znaniecki gathered; there are in any body of voluntarily produced (even for pay) research materials. Typically, they only had one side of the exchange, a common and frustrating condition often bemoaned by biographers and historians. In a commentary on this study, Riley (1963)

> "In all such instances, then, their data refer only to selected members of each group (family) and cover only part of the interaction.

---

[1] Meadow, Greenblatt, Levine and Solomon (1952) report on the use of the index with schizophrenics.

These gaps illustrate an important potential limitation in the use of available data generally: not having been assembled for the purpose of the investigation, the data may be fragmentary or incomplete, thus depriving the researcher of valuable information.

"Another limitation is that such privately owned and spontaneously produced materials may be rare or difficult to obtain. Owners of letters, diaries, or other personal documents may sometimes object to their use for research purposes. . . . Moreover, situations producing appropriate materials may be rare. The continuing exchange of letters, for example, seems to depend upon long-term or frequent separation of the members, as well as upon a custom of detailed letter writing. Nevertheless, there are no doubt many instances in which similar data are available for further research, as, for instance, when servicemen are separated from their families" (p. 242).

Riley thus points out that the dross rate may be high ("situations producing appropriate materials may be rare"), and that population restrictions may be present (". . . exchange . . . seems to depend upon long-term or frequent separation . . . owners may sometimes object to their use"). Specifically for cross-cultural comparison purposes there is the question of differential literacy rates. How many of the Polish peasants could write? If they could not, and had a letter written for them by someone else, say a village scribe, did the presence of this intervening person serve to alter the content of the letters? On the voluntary supplying of the letters, did the correspondent give up only a biased sample? A money incentive might have to be prohibitively high to pry loose some love letters, for example, or letters which detailed complaints about the correspondent's frugality in sending money back home.

In Sunday feature articles, one sometimes reads another group of one-way letters: those sent from children in summer camp to their parents. By themselves, they are instructive of a child's perception of the surrounding world. Salzinger (1958) got the other end as well of such candid correspondence, and content analyzed mail from children and parent--comparing the letters for similarity on "wants," "demands," and "requests."

Janowitz (1958) dealt with letters and diaries captured from German soldiers. His concern was the impact of propaganda on these troops, and "when these letters dealt with the German writer himself, or his small circle of friends, they contained testimony of considerable value. Many made valuable propaganda documents, especially captured undelivered mail" (p. 734). This last point is of interest, for the undelivered mail is a subset of mail which is most recent and most pertinent for propaganda effect evaluation. Letters captured on the person of troops may also be particularly valuable, for they include not only the most recent expressions of feeling, but may also include letters the mailing of which the writer has been uncertain. Such uncertain writings may be prime indicators of attitudes and morale, for the easy stereotypes of "I'm fine ma and the food's not bad" are more likely to be quickly dispatched.

Letters to political figures are another source of data. For some magnificent examples of these, as well as a general treatment of the topic, one can examine Dear FDR (Sussman, 1963, an earlier report of which is in Sussman, 1959. See also Dexter, 1963 a).

A particularly fine discussion of possible error sources in letters is presented in Dexter's (1963 b) chapter on letters to Congressmen. He notes that Congressmen do not necessarily see any cause for alarm in a barrage of negative mail.

> "One 'pro-labor' Senator, out of curiosity, had his staff check up on the writers of 100 letters he received advocating support of a higher minimum wage. It was found that 75 writers were eligible to register, but of these only 33 actually were registered. Furthermore, the letters were advocating his support of a measure on which he had been particularly active, and the content of the mail showed no realization of the stand he had so publicly taken. The Senator could scarcely get excited about these letter writers as either a source of opposition or of support on the basis of that issue" (p. 399).

If the only goal of the Senator were to stay in office, then the mail from an ineligible voter is only so much dross. His greater concern is the lack of any intelligence from the great mass of eligible voters who don't write.

Dexter goes on to discuss why mail is important, but speaks of the congressman's description of "genuine," "junk" and "stimulated" mail. Of interest is the way in which they are discriminated.

> "It (mail) is not believed if "junk," i.e., press releases or other broadcast mailings, nor if it be stimulated. Stimulated mail is not entirely easy to define. In its pure form its consists of virtually identical postcard messages written under the instigation of a single company, union or interest group. (One company even mailed the postcards for its workers, fearing that they would not know who their congressman was.) Congressmen look for signs of stimulation--similarity of phrasing ('They all used the same argument.') or even stationery ('They handed out the paper.') and time of mailing

> ('You could tell the hour or minute someone pushed the button.') . . . it is hard to fool a congressman as to when mail is stimulated. Some organizations urge their members to write in their own words, on their own stationery, and as personally as possible. Congressional assistants tell us that perhaps one in fifty persons who write such a letter will enclose the original printed notice from the organization urging an individualized apparently spontaneous letter" (p. 403).

As for the extent of this false element in spontaneous mail,

> "Most of the mail sent on the Reciprocal Trade Act was in some sense stimulated . . . (for) Eastern and Southern congressmen . . . Westinghouse, Dow, Monsanto, and Pittsburgh Plate Glass may have stimulated 40 per cent or more of all the mail received on the issue in 1954. . . . Mail in favor of reciprocal trade was equally stimulated and perhaps by even fewer prime movers. Our impression is that three-fourths of all antiprotectionist mail was stimulated directly or indirectly by the League of Women Voters" (pp. 403-404).

This selective bias in the population mailing letters to congressmen results in an unvalid generalization on the state of public opinion, but it can serve as evidence on how the major pressure groups within a constituency are responding to a critical bill. The bias itself isn't bad; only not knowing of it is.

Earlier we noted the problem of population restriction bias in suicide notes, observing that less than 25 per cent of all suicides leave final notes. Osgood and Walker (1959) took note of this in their study of motivation and language behavior. Reasoning from behavioral principles, they

9

predicted that the content of suicide notes should differ significantly from control notes and simulated suicide notes. Persons about to take their life should be highly motivated (something of an understatement), and this motivation should increase the dominant responses in their hierarchies; a higher than normal level of stereotypy should be present. Content analysis by six different stereotypy indices supported their prediction. This study is a good example of relating "natural phenomena" that exist in the outside world to principles derived from laboratory experimentation. There are many tests for theoretical postulates available in settings other than the laboratory, and joint testing in the laboratory and outside may yield powerful validity checks.

Spiegel and Neuringer (1963) also tested a specific hypothesis by the employment of suicide notes. They examined the proposition that inhibition of the experience of dread ordinarily evoked by suicidal intention is a necessary condition for suicidal action. They drew on the Los Angeles County Coroner's office for their material, noting a control of "false" suicide notes. Other suicide note studies have been conducted by Shneidman and Farberow (1957) and by Gottschalk and Gleser (1960).

Art work is another expressive personal document that may provide data--as is shown by all the clinical psychologists who look at Van Gogh's paintings and say, "That man was in trouble!" An equally post hoc analysis, but with more analytic elegance, has been contributed by Barry (1957). He studied the complexity of art form as related to severity of socialization. From Whiting and Child's 76 nonliterate societies on which socialization data is available, he found 30 with at least 10 extant examples of graphic art--either displays in museums or illustrations in ethnographic reports. There was a low level association between art form complexity and degree of severity of socialization.

The unknown question is whether a higher or lower level of association would have been detected had the data been available for more than 30 societies. Were those who were more gentle in socialization less likely to produce art work which has survived to the present? Note too that there is the selective screen of museum curators and ethnographers. Materials might have physically survived from the other 46 societies, but have been defined as of insufficient artistic or scientific worth to display behind glass or on paper.

What might be considered an equally primitive art form was studied by Craddick (1962) to infer inhibition. He studied drawings of witches made by children before, on and after Halloween, noticing how large the witches were. The fluctuating size of the witches were attributed to various inhibitory mechanisms. .

A Concluding Note

In this review of archival studies, we have seen the versatility of the written record. Not only has the content of study varied, but also the functions these data have served.

For some research purposes, there were few alternatives to archives--not a particularly luminary recommendation, but certainly a compelling one. What with all the false suicides, for example, there is no choice but to wait until a population operationally defines itself. Once this happens, one can go to farewell notes, biographical material, and interviews with relatives; but one cannot go to the subject. So, too, for the general student of the past. For one like Terman (1917), who chose to study Galton, there was no easy alternative to consulting the written record.[1]

---

[1] In a survey studying intellectual masters, Boring and Boring (1948) treated nonrespondents "as if they were dead: we wrote to persons who were closely associated with them at the time they received their Ph.D.'s. We think these relationships are essentially correct, but they are not incontrovertible. There was, for instance, one delinquent for whom we thought we had surely obtained through collateral correspondence the correct 'master'--until he wrote in late to claim a quite different filiality" (p. 529).

In a limited content area, the archival record provides the dependent variable. Just as votes are the ultimate criterion for the politician, sales and work performance are the ultimate criteria for some applied social scientists. It has been of interest in the history of research in both advertising and personnel that so relatively simple criterion variables have been ignored while less pertinent ones were labored over(measuring "willingness to buy" by questionnaire methods is an example, although it does have some utility in prediction studies).

There are also a few studies in which the records were used as a medium by which theoretical principles could be tested. Such studies are too few, but these records offer superb opportunities to validate hypotheses generated in less natural and more reactivity prone settings. There are restrictions, but it should be recognized that there are restrictions in any single class of information.

> "Skinner and his associates have concentrated on situations in which an animal can perform a particular kind of response repeatedly at a high rate. The findings yielded by this kind of experiment have been extrapolated without much hesitation, and not always with specific empirical warrant, to a diversity of human activities, including those on which the most important social problems hinge" (Berlyne, 1964, pp. 115-116).

Osgood and Walker (1959) used the exterior setting of suicide notes to study the effect of heightened motivation on response hierarchies. Also using records as a testing medium, Mosteller and Wallace (1962) went to records of 1787-1788 for their comparative study of a Bayesian procedure with a classical statistical approach. They demonstrated that both procedures reached the same conclusion on the disputed authorship of some of The Federalist papers.

The great majority of these studies, however, have used the archives for indirect evidence. Stuart's (1963) study of union grievances and the state of race relations, Parker's (1963) study of library withdrawals to show the effect of television, and Craddick's (1962) measurement of the size of witches in a study of inhibition all reveal the inventive unveiling of valuable evidence. But only partial evidence--for the reasons traced in the preceding chapter for care in generalizing from such analyses. Here, as with the running public records, there is a heavy demand for consideration of possible data transformations and for the construction of multiple indices. If it is agreed that the archives typically provide only partial evidence, and if the desirable research strategy is to generate multiple displays of overlapping evidence, then the way in which these partial clues are pieced together is critical.

We should recognize that using the archival records frequently means substituting someone else's selective filter for your own. Although the investigator may not himself contaminate the material, he might learn that the producer or repository already had. A thoughtful consideration of the sources of invalidity may provide intelligence on these, either by suggesting astute hedges or new analyses to answer rival hypotheses. In any event, the Chinese proverb still holds:

"The palest ink is clearer than the best memory."

Chapter 5

Simple Observation

> "Who could he be? He was evidently reserved, and apparently melancholy. Was he a clergyman?--He danced too well. A barrister?--He was not called. He used very fine words, and said a great deal. Could he be a distinguished foreigner come to England for the purpose of describing the country, its manners and customs; and frequenting city balls and public dinners with the view of becoming acquainted with high life, polished etiquette, and English refinement?--No, he had not a foreign accent. Was he a surgeon, a contributer to the magazines, a writer of fashionable novels or an artist?--No: to each and all of these surmises there existed some valid objection.--'Then,' said everybody, 'he must be sombody.'--'I should think he must be,' reasoned Mr. Malderton, with himself, 'because he perceives our superiority, and pays us much attention.'"
>
> Sketches from Boz

Charles Dickens displays a ready touch for observationally scouring the behavior of this mystical gentleman for evidence with which to classify him--even going so far as to put out the hypothesis that the man was a participant observer. In this chapter, the first of two on observation methods, our interest is focussed on situations in which the observer has no control over the behavior or sign in question, and plays an unobserved, passive and non-intrusive role in the research situation. A later chapter details studies in which the observer has played an active role in structuring the situation, but in which he is still unobserved by the actors. "Since we have limited our discussion to measures with low risks of reactivity, the visible "research observer" approach and the participant observation method have been excluded here.

The patently visible observer can produce changes in behavior that diminish the validity of comparisons. Arsenian (1943) noted that the simple presence of an adult sitting near a door seemed to lend assurance to a group of nursery-school children. The opposite change was noted by Polansky et al. (1949) in studying the effect of the presence of observers among young boys at a summer camp. There the observers were a threat and became objects of active aggression. Not only is there change produced which reduces the generalizability of findings, but if one were comparing children in two such settings, internal validity would take a blow.

The effect of the observer may wear off over time, as Deutsch (1949) has shown, and thereby produce a selective contaminant in observational data series. The defense against this is to permit the effect of the observer contaminant to wear off, and start analysis with data subsequent to the time when the effect is negligible. This is similar to experimental controls for practice effects in learning experiments, and presumes that the effect will wear off quickly enough not to waste too much data. And that in turn is based on the researcher's ability to measure the independent effect of observation in the series.

Bales (1950) tested whether different arrangements of observers would selectively bias group behavior. Observers sat with the group, or behind a one-way screen with the group aware they were there, or behind the screen with the group unsure if they were there. He found no difference in group behavior under these conditions. All conditions were applied in a laboratory, however, and all the groups knew they were being tested. These factors might overpower the possibly weaker effects of the physical position of the observer.

No matter how well integrated a participant observer becomes, we feel, he is still an element with potential to substantially bias the production of the critical data. The bias may be a selective one to jeopardize internal validity, or, perhaps more plausibly, it may cripple the ability of the social scientist to generalize his findings very far beyond his sample. A number of writers (cf. Bain, 1960; Gullahorn and Strauss, 1960; Gusfield, 1960, Wax, 1960) have argued for the participant observation method as a device to circumvent some of the contaminations of studies employing an "outside" observer. It may do that, but there is still a high risk of contaminants surviving to invalidate comparisons.

Riley (1963) has suggested that the participant observation studies are subject to two classes of error--"control effect" and "biased-viewpoint effect." The control effect is present when the measurement process itself becomes an agent working for change: "the difficulty with control effect in participant observation, and in many other research designs, is that it is unsystematic . . ." (Riley, 1963 , p. 71). The biased-viewpoint effect includes that we have discussed under the label of intrainstrument processes. The instrument (the human observer) may selectively expose himself to the data, or selectively perceive them, and worse yet, shift over time the calibration of his observation measures.

This has been suggested by Naroll and Naroll (1963), who speak of the anthropologist's tendency to be disposed to "exotic data". The observer is more likely to report on phenomena which are different from those of his own society or subculture than he is to report on phenomena common to both. When the participant observer spends an extended period of time in a foreign culture (a year among the Fulani or six months with a city gang), those elements of the culture which first seemed notable because they were alien may later acquire a more homey quality. His increased familiarity with the culture alters him as an instrument.

White suggests that the control effect may be reduced by the observer assuming an incognito role, even though ethical questions are raised, But,

> "on the other hand, the covert observer may find complete immersion in the system, and subsequent likelihood of a biased viewpoint, more difficult to avoid. Limited to his specified role, he may be cut off from valuable channels of information, unable to solicit information not normally accessible to his role without arousing suspicions" (p. 72).

Associated with this class of observation is the use of the informant, who is a participant observer one selective screen away from the investigator. Back (1960) writes of the traits of the good informant (knowledgeability, physical exposure, effective exposure, perceptual abilities, availability of information, motivation), and points out some of the difficulties of receiving valid and appropriate data from informants. For an earlier treatment, see Campbell (1955).

As a final note on participant observation, we cite Lang and Lang's (1960) report, in which participant observers became participants. Two scientific observers of audience behavior at a Billy Graham Crusade in New York made their "Decision for Christ" and left the fold of observers to walk down the aisle. This is in itself an interesting measure. What a testimony to the Rev. Graham's persuasive skills, when sociological observers are so swayed that they leave their posts. In one sense, it is similar to the research approach followed by Argyle, Gardner and Cioffi (1958) in their study of absenteeism.

Stephen Leacock said, "Let me hear the jokes of a nation and I will tell you what the people are like, how they are getting along, and what is going to happen to them" (Manago, 1962). This may be too haughty a claim for conclusions possible from one set of observational data, but we shall note studies which produce impressive findings from the opportunistic use of observation of events over which the investigator has no control.

These simple observation studies have been organized into the following categories: exterior physical signs, expressive movement, physical location, language behavior (conversation sampling) and time duration. The breadth of these measures is notable, and they are "simple" only in that the investigator does not intervene in the production of the material.

Exterior Physical Signs

Most of the exterior physical signs are durable ones that have been inferred to be expressive of current or past behavior. A smaller number are portable and shorter-lived. The bullfighter's beard is a case in point. Conrad (1958) reports that the bullfighter's beard is longer on the day of the fight than on any other day. There are supporting comments among matadors about this phenomenon, yet can one measure the torero's anxiety by noting the length of his beard? The physical task is rather difficult, but not impossible in this day of sophisticated instrumentation. As in all these uncontrolled measures, one must draw inferences about the criterion behavior. Maybe it wasn't the anxiety at all. Perhaps the bullfighter stands farther away from the razor on the morning of the fight, or he may not have shaved that morning at all (like boxers). And then there is the possible intersubject contaminant that the more affluent matadors are likely to be shaved, while the less prosperous shave themselves.

A less questionable measure is tattoos. Burma (1959) reports on the observation of tattoos among some 900 inmates of three different institutions. The research measure was the proportion of inmates with tattoos: "significantly more delinquents than nondelinquents tattoo themselves." Of course, one could hardly reverse the findings and hold that self-tattooing can be employed as a single measure of delinquency. Returning to the bull ring for a moment, "There are many ordinary bullfighters, but ordinary people do not fight bulls" (Lea, 1949, p. 40).

More formal classification cues are tribal markings and scars. Doob (1961) reports of a walk he and an African companion took through a Nigerian market: "I casually pointed to a dozen men, one after the other, who had facial scars. My African friend in all instances named a society; then he and I politely verified the claim by speaking to the person and asking him to tell us the name of his tribe. In eleven instances out of twelve, he was correct. Certainly, however, he may have been responding simultaneously to other cues in the person's appearance, such as his clothing or his skin color" (p. 83).

In a report whose authors choose to remain anonymous (Anonymoi, 1964), it was suggested that there is a strong association between the methodological disposition of psychologists and the length of their hair. The authors observed the hair length of psychologists attending professional meetings and coded the meetings by the probable appeal to those of different methodological inclinations. Thus, in one example, the length of hair was compared between those who attended an experimental set of papers and those who attended a series on ego identity formation. The results are clear-cut. The "harder headed" psychologists have shorter cut hair than the "softer headed" psychologists. The causal ambiguity of the correlation was clarified when the "dehydration hypothesis" (i.e., that lack of insulation caused the hardheadedness) was rejected by the "bald-head control," i.e., examining the distribution of baldheaded persons (who by the dehydration hypothesis should be most hardheaded of all).

Rosenthal (1962), writing of"the wide variance between private manners and public behavior" of the Japanese, wrote:

> "Professor Enright (British lecturer in Japan) and just about every other foreigner who ever visited Japan have noted with varying degrees of astonishment that there is a direct relationship between the politeness of a Japanese and whether or not he is wearing shoes" (p. 20).

It is quite likely that this relationship reflects the selective distribution of shoes in the Japanese society more so than any causal element, an example of a population restriction. The economically marginal members of the Japanese population should, one would think, be more overt in expressing hostility to foreign visitors than those who are more economically stable--and possession of shoes is more likely linked to affluence than it is to xenophobia.

Shoe styles, not their presence, has been used as the unit of discrimination in the U. S. society where almost everybody does wear shoes. Gearing (1952), in a study of subculture awareness in south Chicago, observed shoe styles, finding certain features of the shoe to correspond with certain patterns of living. In general, the flashier shoe more often belonged to the more culture-bound individual. Similar concern with feet was shown by the OSS Assessment Staff (1948) when, because standard uniforms reduced the number of indicators, special attention was paid to shoes and socks as the sole visible indication "of taste and status."

Despite the general consensus on clothing as an indicator of status, little controlled work has been done on the subject. Flugel (1930) wrote a discursive book on clothing in general, Webb (1957) reported on attitudes and status differentiation among clothing stores, and Kane (1958; 1959; 1962) observed the clothing worn by outpatients to their interviews. Kane has considered pattern, color, texture and amount of clothing--relating these characteristics to various moods, traits and personality changes. Green and Knapp (1959) associated preferences for different types of tartans with need achievement, and it would be of interest to see if this preference pattern were supported in clothing purchased and/or worn.

Jewelry and other ornamental objects can also be clues. Freud gave his inner circle of six, after World War I, rings matching his own. On another intellectual plane, observers have noted that in some societies one can find illiterates who buy only the top of a pen and then clip it to clothing as a suggestion of their writing prowess. One could observe the frequency of such purchases in local stores, or less arduously, consider sales records over time from the manufacturer--considering the ratio of tops to bottoms for different countries or regions. The observation method would have an advantage in that one could make coincidental observations on the appearance of those purchasing the tops alone, or isolate a sample for interviewing. The archival record of top and bottom shipments is infinitely more efficient, but more circumscribed in the content available for study.

As part of their study of the social status of legislators and their voting, MacRae and MacRae (1961) observed the houses lived in by legislators and rated them along the lines suggested by Warner (Warner, Meeker and Eells, 1949). This house rating was part of the over-all social class index produced for each legislator.

Observation of any type of possession could be employed as an index if the investigator knows that there is a clear relationship between possession (ownership) of the object and a second variable. Callouses, for example, can serve as an observable indicator of certain classes of activity. Different sports make selective demands on tissue, for example, and the callouses that result are reliable indicators of whether one is a squash player or a golfer. Some occupations may also be determined by similar physical clues.

With these measures used alone, validity is often tenuous. Phillips (1962) showed a multisource exception in his report of the influence on Miami of the influx of 100,000 Cubans. Two years following the Castro revolution he observed:

. . . bilingual street signs (No Jaywalking; Cruce por la zona para peatones)

. . . "a visitor hears almost as much Spanish as English."

. . . signs in windows saying "se habla espanol"

. . . stores with names like "Mi Botanica" and "Carniceria Latina"

. . . latin-american foods on restaurant menus

. . . supermarkets selling yuca, malanga and platanos

. . . the manufacture of a Cuban type cigarette

. . . radio broadcasts in Spanish

. . . Spanish language editorials in the English language newspapers

. . . service held in Spanish by forty Miami churches

Perhaps Phillips was overstating his case, but the marshalling of so much, and so diverse, observational evidence is persuasive. For a prime source in such studies of the unique character of cities, and their changes, there is that eminent guide--the classified telephone directory. It can yield a wide range of broad content information on the economy, interests and characteristics of a city and its people. Isolating the major U.S. cities, which ones have the highest numbers of palmists per thousand population?

Expressive Movement

The more plastic variables of body movement have interested a large number of investigators, with many of the studies on expressive movement and personality measurement reviewed by Wolff and Precker (1951). Of particular interest in their chapter is the emphasis on consistency among different types of expressive movement. They review the relation between personality and the following measures: facial expression, literary style, artistic style, style of speech, gait, painting and drawing, and handwriting. Not all of these studies are nonreactive, since the central criterion for this is that the subject is not aware of being measured. We refer the reader to this chapter for summaries of earlier literature.

Examples of using expressive movement as a response to a particular stimulus--i.e., stimulus-linked rather than subject-linked--are provided in the work of Maurice Krout (1933; 1935; 1937; 1951; 1954a; 1954b). Although this work was done in a laboratory setting, it was under facade conditions. That is, subjects were unaware of the true purpose of the research, considering the experiment a purely verbal task. There is a good possibility for application of Krout's approach in less reactive settings. He elicited autistic gestures through verbal conflict situations, and his analysis deals primarily with digital-manual responses. An example of his findings is the correlation between an attitude of fear and the gesture of placing hand to nose.

Kinesics as a subject of study is relevant here, although as yet large amounts of data are not available. Birdwhistell (1959; 1960; 1963) has defined kinesics as being concerned with the communicational aspects of learned, patterned body motion behavior. This system of nonverbal communication is felt to be inextricably linked with the verbal, and the aim of such study is to achieve a quantification of the former which can be related to the latter. Some "motion qualifiers" have been identified, such as intensity, range, and velocity. Ruesch and Kees (1956) have presented a combination text-picture treatment in their book, Non-Verbal Communication. An example of the impressionistic style of observation is provided by Murphy and Murphy (1962), who reported on the differences in facial expressions between young and old Russians: "While faces of old people often seemed resigned, tired and sad, generally the children seemed lively, friendly, confident and full of vitality" (p. 12).

Something of the detail possible in such studies is shown in Wolff's (1948; 1951) work on hands. In the first study, she observed the gestures of mental patients at meals and at work, concluding, "I found sufficient evidence that correlations exist (1) between emotional make-up and gesture, (2) between the degree of integration and gesture" (Wolff, 1948, 166). The second study was anthropometric, and she compared features of the handprints of schizophrenics, mental defectives and normals. The hands were divided into three major types: elementary, simple and regressive; motor, fleshy and bony; and sensitive small and large. On the basis of an individual's hand type, measurements, nails, crease lines, and type of skin, she delineates the main characteristics of their personality, intelligence, vitality and temperament.

Without necessarily endorsing her conclusions, we report that she found confused crease-line pattern peculiar to the extreme of mental deficiency and other structural characteristics such as concave primary nails, "appeared to a greater or lesser degree in the hands of mental defectives . . . but were completely absent in the hands of the control cases" (Wolff, 1951, 105).

A journalistic account of the expressive behavior of hands has been given by Gould (1951). Here is Gould's account of Frank Costello's appearance before the Kefauver crime hearings:

> "As he (Costello) sparred with Rudolph Halley, the committee's counsel, the movement of his fingers told their own emotional story.
> When the questions got rough, Costello crumpled a handkerchief in his hands. Or he rubbed his palms together. Or he interlaced his fingers. Or he grasped a half-filled glass of water. Or he beat a silent tattoo on the table top. Or he rolled a little ball of paper between his thumb and index finger. Or he stroked the side piece of his glasses lying on the table. His was video's first ballet of the hands" (p. 1).

Schubert (1959) has suggested that overt personal behavior could be used in the study of judicial behavior. In presenting a psychometric model of the Supreme Court, he suggests that the speech, grimaces and gestures of the judges when hearing oral arguments and when opinions are being delivered, are rich sources of data for students of the Court.

On the other side of the legal fence, witnesses in Hindu courts are reported to give indications of the truth of their statements by the movement of their toes (Krout, 1935). The eminent American legal scholar, J. H. Wigmore, in works on judicial proof and evidence (Wigmore, 1913; 1935) speaks of the importance of peripheral expressive movement as clues to testimony validity.[1]

The superstitious behavior of baseball players has been suggested as an area of study by Sechrest and Wallace (1964). They are interested in such behavior as knocking dust off cleats, amount of preliminary bat swinging, tossing dust into the air, going to the resin bag, and hand wiping on shirts. Their hypothesis is that the extent of such superstitious behavior is related to whether or not the player is in a slump or in the middle of a good streak. This study of superstitious behavior could be extended to other sports in which the central characters are relatively isolated and visible. It should be easier for golfers and basketball players, but were difficult for football players. From a practical point of view, of course, coaches and scouts have long studied the overt behavior of opponents for cues to forthcoming actions. Does the fullback indicate the direction of the play by which hand he puts on the ground? Does the linebacker rest on his heels if he is going to fall back on pass defense? Does the quarterback always look in the direction in which he is going to pass, or does he sometimes look the other way, knowing that the defense is keying on his eyes?

---

[1] As evidence of his own openness to all classes of evidence, Wigmore (1913) reprinted an article of G. Stanley Hall's (1891) on childrens lies.

The eye has been a focus of concern as a potential indicator of interest for centuries. It is a commonly held belief that the eyes "grow big" when looking at a favored object. That there is some truth in this folk belief is shown in two very diverse pieces of evidence. The first concerns the behavior of Chinese jade dealers (Gump, 1962). Over the centuries, the dealers have determined a potential buyer's interest in various stones by observing the dilation of his pupils as the pieces were shown. (Astute buyers counter this by wearing dark glasses.)

The second body of evidence has come from the laboratory. Hess and Polt (1960) measured pupil dilation on 16 mm film and related it to stimulus materials. The stimulus objects were a series of pictures--a baby, mother holding a child, partially nude woman, partially nude man, and a landscape. The six pictures elicited clear-cut differences in pupil size. A similar application of this technique was reported by Foote (1962), based on work under the direction of Hess. Using advertisements as stimulus material, a strong correlation was found between pupil dilation as photographed and advertising ratings as measured by recall scales.

The observation of both exterior personal signs and expressive movement assumes a body of associated knowledge on the meaning of the sign or movement. The issue of confusion of shared symbols comes up often enough within the same society ("Oh, I thought you meant . . .") and is a hazard when doing comparative studies across societies.

> "Professor Ehrenfels pointed out that the Tibetans stick out their tongues to symbolize friendly greeting . . . Ehrenfels also described the initial conditions in which he observed that the Bulgarians indicate assent by a horizontal wagging of the head, not a vertical one. As a traveler asking if indeed this road leads to such a city, his respondents indicated that he was correct both verbally and by head gesture. Had they used only the head gesture, he would not have had the context of similarity in which to anchor the striking conclusion that the head gesture was being used in reverse" (Campbell, 1964, 329).

Or, maybe, the Bulgarians words really meant "no." The proof was found by empirical validation.

Physical Location

The physical position of animals has been a favored measure of laboratory scientists, as well as those in the field. Imanishi (1960), for example, described the social structure of Japanese macaques by reporting on their physical grouping patterns. The macaques arrange themselves in a series of concentric rings, the centermost consisting of dominant males and females.

With people, there are the familiar newspaper accounts of who stood next to whom in Red Square reviewing the May Day parade. The proximity of a politician to the leader is a direct clue of his status in the power hierarchy. His physical position is interpreted as symptomatic of other behavior which gave him the status position befitting someone four men away from the Premier, and descriptive of that current status position. In this more casual journalistic report of observations, one often finds time-series analysis: Mr. B. has been demoted to the end of the dias and Mr. L. has moved up close to the middle.

Feshbach and Feshbach (1963) report on a measure where location was determined by the subjects and not imposed by someone else. The Feshbachs induced fear in ten boys, aged 9 to 12, at a Halloween party by telling them ghost stories. The boys were then called out of the room and administered questionnaires. The induction of the fear state was natural, but their dependent variable measures were potentially reactive. What is of interest to us is a parenthetical statement made by the authors. After describing the ghost story telling situation, the Feshbachs offer evidence for the successful induction of fear:

> "Although the diameter of the circle was about eleven feet at the beginning of the story telling, by the time the last ghost story was completed, it had been spontaneously reduced to approximately three feet" (p. 499).

Somner used the position of chairs in a descriptive way--looking at "the distance for comfortable conversation." Normal subjects were used, but observations were made after the subjects had been on a tour of a large mental hospital. Distances among chairs in a lounge were systematically varied, and the people were brought into the lounge after the tour. They entered by pairs, and each pair was asked to go to a designated area and sit down. A simple record was made of the chairs selected.

The issue here is what one generalizes to. Just as the Feshbachs' subjects drew together during the narration of ghost stories, it would not be unrealistic to expect that normal adults coming off a tour of a mental hospital might also draw closer together than would be the case if they had not been on the tour. Their seating distance before the tour would be an interesting control. Do they huddle more, anticipating worse than will be seen, or less?

Sommer (1959; 1960; 1961) has conducted other studies of distance, and in the 1959 study mentions a "waltz technique" to measure psychological distance. He learned that as he approached people they would back away; when he moved backward during a conversation, the other person moved forward. The physical distance between two conversationalists will also vary systematically by the nationality of the talkers, and there are substantial distances between two Englishmen talking together and two Frenchmen in a conversation. In a cross-cultural study, this would be a response set characteristic to be accounted for.[1]

---

[1] Sommer's work did inspire a study in Germany (Kaminski & Osterkamp, 1962), but unfortunately it is not a replication of Sommer's design. A paper and pencil test was substituted for the actual physical behavior and 48 students were tested in three mock situations: classroom, U-shaped table, park benches.

Other tests include measuring the distance subjects placed photographs away from themselves (Smith, 1958 ; Beloff & Beloff, 1961), and Werner & Wapner's (1953) research on measuring the amount of distance walked under conditions of danger.

Sommer (1960) noted how the physical location of group members influenced interactions. Most communication took place among neighbors, but the corner was the locus of most interaction. Whyte (1950) observed that air conditioners were dispersed in a nonrandom way in a Chicago suburban community, and Howells and Becker (1962) demonstrated that those who sat facing several others during a discussion received more leadership nominations than did those who sat side-by-side.

Leipold's (1963) dissertation carried the work further, paying special attention to the individual response set variable of "personal space," the physical distance an organism customarily places between itself and other organisms. Leipold gathered personality classification data on a group of 90 psychology students, divided them into two groups on the basis of introversion-extraversion and administered stress, praise, or neutral conditions to a third of each group. He evaluated the effect of the conditions, and the tie to introversion-extraversion, by noting which of several available seats were taken by the subjects when they came in for a subsequent interview. The seats varied in the distance from the investigator. In one of his findings, he reports that introverted and high anxious students kept a greater physical distance from the investigator (choosing a farther chair) than did extraverted and low anxious students.

That random assignment doesn't always work is shown in Grusky's work on organizational goals and informal leaders--research conducted in an experimental prison camp. He learned that informal leaders, despite a policy of random bed assignments, were more likely to attain the bottom bunk. Grusky (1959) added such archival measures as number of escapes, general transfers and transfers for poor adjustment to his data. On all of these measures, leaders differed significantly from nonleaders. It must be remembered that this was an experimental prison camp, and there is the risk that a "Hawthorn Effect" may be present. What would be valuable would be another study of regular prison behavior to see if these findings hold in a nonexperimental setting.

The clustering of people has been used as a measure by Spohn and Hunter (1961) who observed patients in a psychiatric ward. Group improvement resulting from various psychotherapy programs was evidenced by the composition and groupings of patients. Gratiot-Alphandery (1951a ,1951b) and Herbinere-Lebert (1951) have both made observations of children during informal film showings. The cluster pattern of children of different age groups was the measure in their work on developmental changes.

On still another plane, the august chambers of the United Nations in New York, Alger (1963) observed representatives at the General Assembly. Sitting with a press card in the gallery, he recorded records of 3,322 interactions among representatives at sessions of the Administrative and Budgetary committee. Each interaction was coded for location, initiator, presence and/or exchange of documents, apparent humor, duration, etc. His interest was in defining the clusters of nations who typically interacted in the committee.

Using the same approach, it may be possible to get partial evidence on which nations are perceived as critical and uncertain during debate on a proposed piece of UN action. Could one define the marginal, "swing" countries by noting which ones were visited by both Western and Bloc countries during the course of the debate? Weak evidence, to be sure, for there is the heavy problem of spatial restriction. One can only observe in public places, and even expanding the investigation to lobbies, lounges and other public meeting areas may exclude the locus of the truly critical interactions. This bias might be selective, for if an issue suddenly appeared without warning, the public areas might be a more solid sampling base than they would be for issues which have long been anticipated and which could be lobbied in private. That the outside observer must have a broad understanding of the phenomenon and parties he is observing is indicated in Alger's study. He comments on the high level of interaction with the Irish delegate, which was not a reflection of the political power of Ireland, but instead the result of the easy affability of that delegate. This affability might truly influence the power position of his country, and hence be an important datum in that sense, but it is more likely to confound comparisons if it is used as evidence on a nation.

Up to now, all the observation has been of people--with only a few chimpanzees and air conditioners thrown in. One may also observe things, aspects of which are more fully covered in our chapter on trace and erosion methods.

Mechanics in an automotive service department are the data gatherers in another study of audience measurement. A Chicago auto dealer (Anonymous, 1962) estimates the listenership of different radio stations by having his mechanics record the frequency of the station last listened to by the customer. The dealer, Z Frank, checks more than 50,000 dials a year with less than 20 per cent duplication of dials. This approach provides data which the dealer employs in selection of radio stations to select to carry his advertising. The sampling is valid if (1) the goal of the radio propaganda is to reach the same type of audience which now comes to the dealership and (2) if a significant number of cars do not have radios. If the latter case were true, that many of the cars are without radios, then a partial and possibly biased estimate of the universe is obtained. In a validation note, it is reported, "We find a high degree of correlation between what the rating people report and our own dial setting research" (p. 83). Neither one of these measures could be considered as a definitive validator, and we have a case of egalitarian mutual validation.

The same approach could be used to study the selective appeal of different radio stations. Knowing that various shopping centers draw customers from quite discrete economic populations, one could observe dial settings in cars parked in the shopping centers and compare them. As a validation check on the discrimination among the centers, one could (in metropolitan areas) note local tax stickers affixed to the automobiles and compare these to the economic data reported for tax areas by the U. S. Census.

Similarly, it would be possible to measure the range of advertising effectiveness by determining the neighborhoods from which customers come in terms of distance from the center. Changes in the range of advertising due to promotional schemes could be measured by a time comparison, again using the tax stickers as indicators.

Barch, Trumbo and Nangle (1957) used not the source of automobiles, but their behavior in a study of conformity to legal requirements. We are not sure if this is more properly coded under "expressive movement," but the "physical position" category seems more appropriate. They were interested in the degree to which turn signalling was related to the turn signalling behavior of a preceding car. For four weeks they recorded this information:

1. Presence or absence of a turn signal
2. Direction of turn
3. Presence of another motor vehicle 100 feet or less behind the turning motor vehicle when it begins to turn
4. Sex of drivers

Observers stood near the side of the road and were not easily visible to the motorists. There was the interesting finding that conforming behavior, as defined by signalling or not, varied with the direction of the turn. Moreover a sex difference was noted. There was a strong positive correlation if model and follower were females, and also a high positive correlation if left turns were signalled. But on right turns, the correlation was low and positive. Why there is a high correlation for left turns and a low one for right turns is equivocal. The data, like so many simple observational data, don't offer the "why." To better infer that, one needs a degree of experimental control usually not present.

Observation of Language Behavior: Conversation Sampling

Language is a hoary subject for observation, with everything from phonemes to profanity legitimate game. Our interest here is more circumscribed and centers on language samples collected unobtrusively, and in a nonreactive manner. This means excluding such useful research as Mahl's (1956) study of patients' speech in psychotherapy sessions. The incidence of stuttering, slips of tongue[1] etc., is important data, because the data was collected in a therapist-patient setting, it does not apply here. Instead, we have taken one area of language research, conversational sampling, and traced it historically to illustrate the methodological issues.

Dittmann and Wynne (1961) demonstrate a modern approach. They coded verbal behavior, with the language source a radio program--the NBC show "Conversation." To study emotional expression, the authors examined "linguistic" phenomena (junctures, stress, pitch) and "paralinguistic" phenomena (voice set, voice quality and vocalizations of three types). A problem comes from the possibility that a man's awareness of participation in a radio show--particularly the effects of nervousness on speech--could lead to conditions that bias the production of the critical responses.[1] More satisfactory is language analysis which draws its samples from speech of subjects unaware of observation. One of the earliest mentions of conversation as a source of psychological data subject to quantification is made by Tarde (1901). Although he himself performed no studies on conversation, Tarde made several suggestions for potential areas of study, such as speed of talking as a function of the particular culture, and categorization of topics as a function of social class differences.

---

[1] We would be curious to read the findings of a study which investigated slips of the typewriter as a measure. The employment of these regularly appearing slips somehow evaded Freud (1920) in his major work on the topic.

For the first reported study of conversations, we can look at H. T. Moore's (1922) work on sex differences in conversation--a canny and delightful research that triggered a whole series of hidden observer language studies.

Moore sought to prove that there was a definite mental differentiation between the sexes, regardless of what previous studies (to 1922) had shown. To test this, he argued for a content analysis of "easy conversation." Especially at the day's end, he held, conversation should provide significant clues to personal interest.

So Moore slowly walked up Broadway from 33rd Street to 55th Street about 7:30 every night for several weeks. He jotted down every bit of audible conversation and eventually collected 174 fragments. Each was coded by the sex of the speaker and whether the company was mixed or of the same sex. It is not necessary to cite his findings at length, but one should not pass attention: in male to male conversations Moore found 8 per cent in the category "persons of opposite sex;" for female to female conversations, this topic occupied 44 per cent of the language specimens.

---

[1] Kramer (1963) has reviewed the literature on the nonverbal characteristics of speech, concentrating on personal characteristic and emotional correlates. In a later article,(Kramer, 1964), he reports a methodological study of methods of eliminating verbal cues. The three major methods are: a constant ambiguous set of words for various emotional expressions; filtering out the frequencies which permit word recognition; speech in a language unknown to the listener.

Some of the limitations of conversation sampling are obvious. Moore could record only intelligible audible conversation. Speech that is muttered, mumbled or whispered may contain significantly different content than loud and clear speech. The representative character of the speech samples is further questioned by the representativeness of the speakers. Walkers on Broadway are probably not even a good sample of Manhattan. In short, there is a strong risk of sampling rigidity in both the talkers and the talk.

We can look, chronologically, at the conversation sampling studies that followed Moore's and note the efforts of other investigators to reduce error due to data collecting procedure.

Landis and Burtt (1924) published the first study stimulated by Moore's classic. They were sensitive to positional sampling biases and improved upon Moore's procedure by sampling a wider variety of places and situations. With an experimenter who "wore rubber heels and cultivated an unobtrusive manner," they gathered samples of conversation, adding an estimate of the social status of the speaker. The broadened locations included street cars, campuses, railroad stations, athletic events, parties, department stores, theatre and hotel lobbies, restaurants, barber shops, churches, and streets in both commercial and residential areas. After their analysis, Landis and Burtt concluded that the source of the collection was only a minor factor. Landis (1927) broadened the sampling base even further, reporting in an article titled "National Differences in Conversation." He sampled conversations in Oxford and Regent Streets in London and compared these results to the earlier Landis and Burtt (1924) findings from Columbus, Ohio.

The monitoring of telephone conversations was the device by which French, Carter and Koenig (1930) measured the degree to which the most common words contributed to the total word usage of conversations. This study of the repetitiousness of language was later used as a control for the repetitiousness of speech of schizophrenics and college students (Fairbanks, 1944).

Stoke and West (1931) tried to limit the number of variables in their conversational sampling study and restricted the sample to undergraduate college students, sampling from random bull sessions held at night in residence halls. The unknown observers were 36 college students, participant observers, who worked with a checklist of probable topics and the date and the number of people in the conversation. A limitation of this "observe--withdraw--record" approach is that the observer can not hope to record adequately the duration of responses. Stoke and West eliminated this component from their data collection--an example of rigidity of the instrument. The approach is also vulnerable to the criticism that the observers' reports are subject to bias, beyond the initial selective perception one, because of the gap between event and notation.

Moving away from the campus, and more to the Moore approach, Sleeper (1931) sampled conversations in the upper level of Grand Central Terminal in New York, during the rush hour from 5:00 p.m. to 6:30 p.m. Sleeper's procedure reflected the dross rate problems of such data as he added a recording variant by excluding all "environmentally stimulated" conversation.

McCarthy (1929) and Mabie (1931) employed free play periods to sample the conversation of children. The visibility of the recorder is a strong problem here and the studies represent reactive methodology. It may be that, as Mabie claims, the children's awareness of her presence had no effect on their conversation. This is uncertain, and our inclination would be to consider that her presence, notebook in hand, would introduce a strong risk of biasing the character of the overheard statements. When asked by the children what she was doing, Mabie told them that she was writing down what the children liked to do during play period. That response itself could predispose the children to verbalize more frequently evaluative type comments.

Surreptitious observation is the only class which fits into what we would call nonreactive testing. Take the studied surreptitiousness of Henle and Hubble (1938). Students were again the subjects, and

> "The investigators took special precautions to keep subjects ignorant of the fact that their remarks were being recorded. To this end, they concealed themselves under beds in students' rooms where tea parties were being held, eavesdropped in dormitory smoking rooms and washrooms, and listened to telephone conversations" (p. 230).

Without extending their explanation, Henle and Hubble report that "unwitting subjects were pursued in the streets, in department stores, and in the home."

Escaping out from under the bed, Carlson, Cook and Stromberg (1936) studied sex differences in conversation by monitoring lobby conversations at the intermissions of 13 regular concerts of the Minneapolis Symphony and at six University of Minnesota concerts. The self selection of subjects may be a serious risk to external validity (who goes to Minneapolis Symphony concerts and who doesn't?). The whispering problem is not so great in a research setting like a crowded theatre, where a premium is placed on loudness.

The size of the group is a clear influence on the degree to which the experimenter must mask himself. For observing two person communication, it well may be necessary to secret oneself under a bed. In a large public gathering, the problem of visibility is solved; the individual providing the conversation sample expects to find unfamiliar people close to him, and the experimenter need not hide. He need only hide his recording of the language. But it isn't so simple as that. Even though the presence of the observer may cause no surprise, the same situation which permits acceptance of the stranger may also have worked to inhibit a class or classes of verbal behavior. For some experiments this may be unimportant--those in which the difference between public and private utterances is negligible. For other experiments, it may be substantial. This is an empirical question for each experimenter to solve. It must be accepted as one of the possible content limitations of conversational sampling.

Watson, Breed and Posman (1948) displayed their concern about the representativeness of college students by deliberately excluding them from their sample of New York talkers. No campus locations were used, and an attempt was made to eliminate "anyone distinguishable as a college student." Working at all times of day and night, they sampled uptown, midtown and downtown Manhattan, including the following locales: business, amusement and residential streets and parks; subways, busses, ferries, taxies and railroad stations; lobbies of movie houses and hotels; stores, restaurants, bars, night clubs. Each observer recorded verbatim what he had heard as soon as possible after hearing it. The sampling of respondents was resolved as well by Watson, Breed and Posman as it has been by anyone.

Contrast this with the participant observation type of approach suggested by Perrine and Wessman (1954). The investigator posed as a stranger to the state, initiated casual conversation with subjects and then directed conversation to political issues by commenting on recent newspaper headlines, etc. The conversation was recorded as soon as possible after leaving the subject, along with sex, race and location, and estimated age and socioeconomic class.

The enormous methodological issue in this type of conversation recording is the 60 to 70 per cent refusal rate--contacts who did not provide the desired information. If nothing else, the eavesdropping approach reduces problem of self selection of the sample, at least that bias attributable to willingness to participate in a survey. Not everyone will chip into a conversation with some stranger who wants to talk about state politics. To use such data as the basis of inferring the state of public opinion is dubious.

Doob (1961) in a recent report talks of a girl in an African market who "was carefully shadowed in the interest of scholarly research." In an approach described by Doob as "unsystematic eavesdropping," he notes that

> "She began talking, and listening, before she entered the market's gate. Within a period of ten minutes--the duration of the research--she spoke with more than twenty people: some she greeted perfunctorily, others she talked to for a few moments concerning relatives and friends. No political or cosmic thoughts were aired" (p. 144).

Of interest is his point on a possible ethnocentric bias among foreign observers in Africa:

> ". . . whereas people in the West . . . are likely to keep themselves occupied and to avoid long periods of complete solitude or, in contact with others, of silence, it may be that many Africans are perfectly content to be unoccupied except by their own feelings and thoughts and sense of well being" (p. 144)

All of the studies of conversation reported here have relied on a content analysis of the conversational samples gathered. The essential problems have been the representativeness of the sample collected. The unobserved observer (secreted under a bed or among a crowd), must be sensitive to the limitations of self-selection of subjects, a problem of external validity, and the limitations of the probable partial character of public conversation samples. Any public conversation may be constrained because of the "danger" of being overheard. Many of the inaudible comments in public are likely to be drawn from a different population of topics than those loudly registered. Moreover, as we noted earlier, the method requires a careful selection of both place and time sampling units to increase representativeness, and these controls will not be the same over different geographic locales. Sampling bus conversations in Los Angeles and in Chicago yields a population of very different subjects. Moreover, these data are typically loosely packed, and it takes a substantial investment in time and labor to produce a large enough residual pool of relevant data. For all these limitations, however, there are research problems where private commentary is not a significant worry, where the adroit selection of locales and time can circumvent selective population characteristics, and where the issue is of sufficient currency in the public mind to reduce the dross rate. For these situations, conversational sampling is a sensitive and faithful source of information.

Time Duration

The amount of attention paid by a person to an object has long been the source of inferences on interest. For research on infrahuman species, notoriously incompetent at filling out interest questionnaires, visual fixation has been a popular research variable.[1] With humans, who can fill out questionnaires, time duration measures have been less popular, but are not uncommon.

---

[1] See the recent study of Berkson and Fitz-Gerald (1963) on the effect of introducing novel stimuli into the visual world of infant chimps.

Frequently, a duration variable is imbedded in a body of other measures. Moore (1917) measured anger, fear and sex interests by giving a subject multiplication tasks and then exposing him to distraction of different types. The time taken to complete the tasks was taken as the measure of interest: the longer the time, the greater the interest in the distracting content.

Melton, in his study of museum visitors, hypothesized a positive relationship between the degree of interest shown by a visitor in the exhibits and the number and quality of the permanent educational results of the museum visit (Melton, 1935). Melton was very careful to study response set biases and situational cues which would contaminate his measure of the duration of observed time spent in viewing an exhibit. He demonstrated the "right turn" bias, and experimented with changing the number of exits and installing directional arrows--all elements which significantly affected the length of a visit. In one finding, for example, he reports that the closer an exhibit is to an exit, the less time will be spent at it. He posits an "exit gradient." Going further, he talks of the number of paintings in a gallery, the degree of applied or fine arts in the room, and comes up with findings on "museum satiation." As the number of paintings in a gallery increase, the average total time in the gallery also increases, the total number of paintings visited increases, and the time per painting visited does not decrease but increases. Melton's attention to these cues provides a model seldom followed in simple observational research.

Washburne (1928) reported on an experiment conducted in the Russian school system, which conceivably could have used time for a measure. Each child in the school was given his own garden plot, and at the same time had joint responsibility in a common garden tended by all the children. It is reported that records were kept to show the relative amount of interest that each child had in the two types of work. While no mention is made of the measure used to determine amount of interest, it would seem that time might be one appropriate measure. Because it can be assumed that for equal care a greater amount of time would have to be spent on the individual garden, adjustments would have to be made in comparing the times for the two gardens.

One of the small group laboratory studies Exline (1962) had observers behind one-way mirrors who noted visual fixation duration and direction. His measures could be used in naturalistic observation research. Among them were:

1. Time spent looking at someone while speaking to him.
2. Time spent looking when being spoken to.
3. Time spent looking when target is talking to third party.
4. An index of mutual glances.

The lack of general emphasis on time duration methods is partially due to difficulty of measurement. For accurate observation, the hurly-burly conditions of a natural setting are damaging; the laboratory control over instrumentation is almost necessary if precise observations of small time units are to be reached.

Sometimes this can be circumvented by a measure in which time is scaled in grosser units than microseconds. Jacques (1956; 1957) for example, defined "responsibility" by measuring how long a worker is allowed to commit the resources of his task without direct supervision. Observation yields "a time span of responsibility" and a descriptive measure of the worker. For a duration measure like this, it would be foolish to calibrate the measurement units so fine as seconds. Many researches demand ultrafine discrimination of time, and for them, natural observation is an awkward method. But where the unit is broader, observation in a natural setting becomes both feasible and desirable. Sometimes it is enough to say, "Professor X's interest in cutaneous sensation extended over a career of 38 years."

Time Sampling and Observation

For the permanent physical cues of observation--the scars, tattoos, and houses owned--the timing of when an observation is made may be relatively unimportant. It may be possible to conjure up conditions in which a tattoo may be so placed that it is differentially visible at various points in a day (with or without jacket, for example), but for the most part, the exterior signs are quite invulnerable to time linked variance.

Many of the other simple observation materials--expressive movement, physical location and language--are, however, subject to the objection that the critical behavior is variable over a day or some longer time period. The risk, of course, is that the timing of the data collection may be such that a selective population periodically appears before the observer, while another population, equally periodically, engages in the same behavior, but comes along only when the observer is absent. Similarly, the individuals's behavior may shift as the hours or days of the week change. The best defense against this source of invalidity is to randomly sample time units.

Working in an industrial setting, Shepard and Blake (1962) observed employees and judged whether they were working or not. By a time sampling design, they found a strong decline in percentage of workers working between 10:30 and 11:00--the time of a daily supervisor meeting which drew them away from direct control over employees.

> "Hence, the composition of supervisors' meetings was changed so as to ensure continuous supervision in the shop . . . thus the managers are correct in their assessment of the workers as lazy and

and irresponsible, and in their conclusion that more consistent control and direction are needed to correct for their tendency to be irresponsible." (p. 88-89)

The technique has been extensively used in nursery settings. where there is a particular need for it because of the greater periodicity of behavior of infants and young children. Arrington (1943) has pointed out many of the factors which must be considered in assessing the results of time sampling recorded by the observer. The length of the time sample must be chosen in accordance with behavior to be observed--if acts of aggression involving preliminary acts of aggregation is the primary interest, samples must obviously be longer than one minute. In the nursery setting, time sampling can most nearly overcome the factor of "observation consciousness," which Arrington feels is negligible at early ages. Degree of sophistication, familiarity with the observer, previous experience in being observed, type of situation, and number of individuals in the situation are also thought to be factors contributing to "observation consciousness."

One of the important time sampling studies of observation in a nursery setting was conducted by Thomas (1929). In recording the activities of the children, Thomas made use of a mapped floor plan and coded movement. Olson (1929) used similar procedures, but concentrated on oral habits rather than movement patterns.

Barker and Wright (1951; 1954) have adopted an opposite strategy to time sampling. They sought to avoid the problem of selective behavior over time by a saturation method, Rather than sample behavior, they censored it. In their 1951 study, observations were made of one child for an entire day, with minute by minute notations. Eight observers were used alternately, each being wholly familiar with the child. For any child under ten, the authors feel the effects of observation are negligible. This may be subject to doubt, however, particularly in view of recorded statements detailing interaction between the observer and the child.

This strategy does solve a problem, but it provides other ones. It is practical for only relatively short periods of time (imagine following a boy for a year), and the method is predisposed to measurement of individuals, not groups. This latter point may be important for the probability of reactive effects creeping in, for as we noted above, the size of the group may be an important factor in the degree of observation consciousness. A person tailing you about all day is quite different from the one next to you in a theater lobby.

Yet these limitations are no more punishing than the limitations of other approaches, and the subtlety of links between behaviors can hardly be better described.[1] Either way, sampling or censusing, a measure of control is added over a usually uncontrolled variable.

Over-All Comments on Simple Observation

The emphasis of this chapter has been on research in which the observer is unobserved, and in settings where the investigator has had no part in structuring the situation. The secretive nature of the observer, whether hidden in a crowd or miles away before a television screen, protects the research from some of the reactive validity threats. The subject is not aware of being tested, there is thereby no concomitant role playing associated with awareness, the measurement does not work as an agent of change and the interviewer (observer) effects are not an issue.

---

[1] Edmond de Goncourt, wrote of the goal of the Goncourt Journal: "What we have tried to do, then, is to bring our contemporaries to life for posterity in a speaking likeness, by means of the vivid stenography of a conversation, the physiological spontaneity of a gesture, those little signs of emotion that reveal a personality, those imponderabilia that render the intensity of existence, and, last of all, a touch of that fever which is the mark of the heady life of Paris. . . ."

Moreover, there is the great gain that comes from getting the data first hand. In studies of archival records and in the examination of trace and erosion evidence, there is always the uncertainty that others who came between the data and the investigator, processing or pawing it, left their own indistinguishable marks.

The first-hand collection of the data, usually of a contemporaneous event, also allows the gathering of other information to reduce alternative hypotheses. One may note characteristics of the subjects which permit a testing of rival hypotheses about the selective composition of the sample. To be sure, these are limited to visual cues mostly, but they can be extremely helpful. Similarly, the ability to observe the subjects in the act permits one to designate the individual actors--either for follow-up observation, or for study with other instruments like the questionnaire or interview. Such follow-up of individuals is difficult or impossible with archival and trace measures.

It would be difficult to underestimate the value of this potential for follow-up. One of the singular gains of simple observational studies is that it is a procedure which allows opportunistic sampling of important phenomena. Because it is often opportunistic, there is the attendant risk that the population under observation is an atypical group, one unworthy to produce generalizations. The follow-up studies may protect against this risk, as can adroit use of locational and time sampling.

Against this impressive list of gains must be balanced some possible sources of loss. Prime among these is the danger that the data gathering instrument, the human observer, will be variable over the course of his observations. He may get less conscientious as boredom sets in, or he may become more attentive as he learns the task and becomes involved with it. If there are any ambiguities on how the behavior is to be coded, the effect of time may be to reduce the variation of coding (increase intra-observer reliability) as he works out operating definitions of which behaviors go with which codes. All of these can work to produce spurious differences in comparisons. That this may apply to the principal investigator, as well as his aides, has been demonstrated by Rosenthal (1963) who has conducted a fascinating set of experiments on "the social psychology of the psychological experiment." Rosenthal's work places a greater emphasis on the necessity for saturated training of observers, hopefully under a "blind" condition in which they do not know what a "good" result or behavior will be.

The one other significant issue under the label of reactive threats comes from response sets.on those observed. To a large measure, these are knowable--either through application of research conducted in other setting, or through direct observation of behavior under different conditions. Hopefully, there will be enough variation in the settings available for sampling to examine whether any systematic response sets are at work, and which can be isolated from other possible variance sources. This is awkward when one is not actively manipulating the environment, and becomes one of the strong arguments for the unobserved observer to surrepticiously and systematically alter the research environment in an undetectable way.

The populations available for observation fluctuate according to both time and location. Thus, some caution should be employed in generalizing from research which gathered observation at one time in one place. If generalizations about the subject matter of conversations are to all people, and content varies by age, then the "place" should be considered as a sampling universe including manifold locations which are likely to draw on different populations. When the concern is generalizing to more limited settings, say a study of the effect of different treatment conditions in prisons, then the place sample should be more than one prison for each condition hypothesized to have an effect.

It is not always possible to draw elaborate locational samples, but that should not deter observational research. If the setting is circumscribed by practical conditions, a proper defense is to employ time sampling methods. Limited to a population of "tour" visitors to a mental hospital, one must bear the cross of a self-selected population unlikely to be representative of too much. Imposing a time sampling design, observing different groups who come in different days or different months, for example, would markedly improve the solidity of a shaky base.

Both time and locational sampling should be employed if possible, for both empirical research and introspection suggest that population variation is a substantial issue in observation. An added gain is that the investigator can also vary his observers over the sampling units and randomly assign them to different times and locations, thus adding a badly needed control. Not all research possibilities afford this chance, but it is a goal to be reached for.

The same time and locational sampling strategy will also help to counteract some of the risks in selective content. The population varies over place and time, and the content of their behavior similarly varies. If one can broaden the sampling base, he can expand character of material available for study.

It is not possible to know all about college students if observations are limited to afternoons in the fall; when these observations fall on Saturday, worse yet.

All the finesse of the skillful investigator will not solve some content limitations. Much of behavior is precluded from public display and is available only through unethical action, elaborate instrumentation or some titantic combination of both. This is potentially a variable problem across cultures, as one notes members of certain societies willing to display classes of behavior that are hidden or taboo in others. Cross-cultural observational studies are thereby threatened not only by the ethnocentric attribution of the meaning of acts (the Tibetans sticking out tongues), but also the frequency of holes in observable behavior across societies. As one increases the number of societies, of course, the greater is the probability that an incomplete set of observations of public behavior will be available over all.

Finally, the advantage of being on the scene means that often there is a substantial amount of behavior produced about which the investigator has no interest. Because one can't often predict when a critical event will be produced, it is necessary to wait around, observe, and complain about the high dross rate of such a procedure. The payoff is often high, as in the case of one patient observer who knew critical signs and was immortalized in the song, "My Lover was a Logger." The waitress sings,

"I can tell that you're a logger,
And not just a common bum,
'Cause nobody but a logger
Stirs his coffee with his thumb."

Chapter 6

Contrived Observation: Hidden Hardware and Control

This chapter speaks of the investigator intervention into the observational setting. In the simple observational studies, research is often handicapped by the weaknesses of the human observer; by the unavailability of certain content and by a morass of uncontrolled variables which threaten the validity of comparisons. To reduce these threats, a number of investigators have opted to actively vary the setting or to substitute hardware devices for human record keeping observers. We avoid here examples of the "Speak clearly into the microphone, please" approach. The emphasis is on those investigations in which the scientist's intervention is not detectable by the subject and where the naturalness of the situation is not violated.

Hardware: Avoiding Human Instrument Error

When the human observer is the recording agent, all the fallabilities of the organism operate to introduce extraneous variance into the data. People are low fidelity observational instruments. We have already noted how variable can be recording and interpretation over time, as the observer learns and responds to the research phenomena he observes.

The fluctuations of this instrument can be brought under some degree of control by random assignment of observers to locations and time units. Random assignment will not, however, create a capacity in the organism which is not there, nor eliminate response sets that are characteristic of all members of the society or subcultures from which the observers are drawn.

Osgood (1953) illustrates the capacity weakness of the human observer in his comments on studies of language behavior in the first four or five months of a human's life.

> ". . . from the total splurge of sounds made by an actively vocal infant, only the small sample that happens to strike the observer is recorded at all" (p. 684).

Not all observable behavior is so complex or so rapid, but enough is to cause the consideration of a substitute mechanism for the observer. It might not be so bad if there were a random loss of material when the observer's perceptual system got overloaded. Unhappily, the nature of the material noted and not noted is likely to be a function of both the individual's idiosyncracies and the systematic response sets learned in a given society. Again speaking of speech studies of early childhood, Osgood comments,

> "The inadequate recording methods employed in most of the early studies make the data of dubious validity. The typical procedure was merely to listen to the spontaneous vocalizations of an infant and write down what was heard. The selective factor of auditory perception--listeners 'hear' most readily those sounds that correspond to the phonemes of their own language--was not considered" (p. 684).

Those same biases are at work with the recording of any language system that is unfamiliar to the observer--whether it be the occult language of a child, or the unfamiliar tones of a foreign language. Webb (no date) has noted this in his study of orthographies in African languages. The basis of his analysis were written records of the languages, many of which had been produced by missionaries, explorers and other foreign nationals who came to Africa and learned the indigenous languages. In transcribing the sounds of these languages for others, there were selective approximations of the true sound, influenced by the tonal pattern of the characters in the observer's native language. Thus, German observers heard umlauts that evaded the British. There is some possibility of control over this particular bias, since for some of the languages there are written transcriptions of the same words by nationals of various countries; these can be matched against the known tonal characteristics of the European languages to correct for the selective hearing. When multinational observations are missing, the task is much more difficult, and one must make inferences about the effect of response sets learned from the observer's native language on his interpretation of the indigenous African language.

Other methods of indexing the frequency distribution of characters are available. Commercial manufacturers sell type in fonts which contain unequal numbers of sticks of type for the different characters. At some time in the past, the distribution estimates were determined. The determination may have come from an empirical study of distributions, or it may have been merely a guess. Learning which allows one the chance to plug into other research outside the linguistic journals. For two of the seven languages studied by Webb, such distributions were available, both of which had been the product of careful investigation, and both of which gave an independent test of his sampling of language specimens. Sales records, if available, could also be used. What is the sales incidence for replacement characters? These data can be considered in consort with the unequal initial distribution of characters in the font as a check on the accuracy of that initial distribution. The measure is bounded by the assumption that all pieces of type are manufactured from the same material. There is the possibility that the producer, knowing of the differential use of characters, made the more common letters from a different alloy than the rarer ones. This would contaminate the sales measure, but knowledge of the error source would lead to another measurement (the metallurgical composition) in its stead. Finally, one may observe the differential wear on characters in use, correcting for the different numbers of pieces of type for each character. This method is a tertiary approach and demands that all materials started with an equal or known degree of wear. It is less applicable for typewriters, since there is an interaction between strength of fingers and individual characters. The "i" on a standard typewriter keyboard is depressed by the strong right middle finger, while the "a" is depressed by the weak left little finger. An electric typewriter solves this issue by making relatively constant the force with which all characters strike the typewriter roller.

A major gain from hardware recording, of course, is the permanence of this complete record. It is not subject to selective decay, and can provide the stuff of reliability checks. Further, the same content can be the base for new hypothesis testing not considered at the time the data was collected. Or material that was originally viewed as dross may become prime ore.

Hardware, of varying degrees of flexibility, has been used throughout the history of scientific observation. To reduce the risk of forgetting, if nothing else, permanent records were made of observed behavior. They may have noted less than the total behavior, but they did serve to reduce reliance on human memory. Boring (1961) talks of Galton, "an indefatigable measurer."

> "He used to carry a paper cross and a little needle point, arranged so that he could punch holes in the paper to keep count of whatever he was at that time observing. A hole at the head of the cross meant greater, on the arm equal, and on the bottom, less" (p. 154).

Evolving from such a paper and pencil recording method was the constriction of communication developed for work in small group research in laboratory settings. Artificially, the participants were (or are) required to limit all communications to written notes, saved by the investigator to provide a full record of all communication among participants. This is a very low cost device, much cheaper than tape recording, but its stilted quality suggests a very high risk price for a very low dollar cost.

Other aids to the observer are pieces of apparatus which allow him to more quickly or more thoroughly record his observations. Sometimes the gain comes from forcing the observer into utilization of a series of varied codes, sometimes it is just the gain of having a more permanent record of his perceptions of the behavior under study. Steiner and Field (1960), for example, timed vocal contributions to a group discussion by means of a polygraph, and a popular supplementary device has been Chapple's Interaction Chronograph, a recent use of which is reported by Chapple (1962).

A big boom exists, and properly, for audio tape recorders. With the development of superior omnidirectional and highly directional microphones, many of the former mechanical limitations have been resolved. The tape becomes the first source of data, and it is often considered as the initial input into a hardware system. Thus, Andrew (1963) took tapes of sound patterns from primates (including man) and fed them into a spectograph for more detailed analyses.

Similarly, Heusler, Ulett and their associates(Heusler et al., 1959; Heusler et al., 1960; Callahan et al., 1960; Ulett et al., 1961a, 1961b) developed what they termed a noise level index for hospital wards. Their substantive concern was measuring the effects on drugs on hospital patients, and they planted tape recorders to pick up ward noises. These sounds are meshed in an integrator which provides a numerical total of the activity. Originally, this noise level had been rated by judges; in later work, however, the authors used a direct index of noise level, thus reducing confounds, among them the possible confounding due to a judge's recognition of a patient's voice.

A highly opportunistic use of audio tapes was demonstrated by Matarazzo and his associates (Matarazzo et al., 1964) in their study of the effect of speech duration. The National Aeronautics and Space Administration made available to these investigators the audio tapes of conversations between the astronaut and his ground communicators for two orbital flights. From these tapes, and the published transcripts of the communications (NASA Manned Spacecraft Center, 1962a; 1962b) they coded the duration of each unit of speech by means of an Interaction Chronograph. These data provided a test of propositions developed in the experimental laboratories and previously reported (Matarazzo, 1962a; Matarazzo et al., 1963). It could hardly be claimed that the astronauts were thinking of Matarazzo's hypotheses at the time they were steering their craft, and the astronaut findings supported the work of the laboratory. This highly imaginative research dipped into archives that were available, archives known to not suffer from intermediary distortions.

The fidelity and content breadth of the audio tapes give them an edge over written records for archival analysis. Not only are they uncontaminated by other hands, but they contain more pertinent material physically unavailable on the written record. Matarazzo, for example, could not have conducted so accurate a study if he were limited to transcripts alone. Interested in duration of speech, where the natural unit is a second of time, he would have had to make estimates of duration from word counts, which are pockmarked with substantial individual response set errors in rate of speech, different levels of noncontent interjections (um-mmm's) and such.

There is a weighty mass of research material almost untouched by social scientists, although known and used by historians. It is found in the oral archives of the national radio and television networks, who have kept disc and tape recordings of radio and television shows over the years. The recent advent of video tape recordings has provided another dimension to these archives, and the recordings are available for all with the inclination to use them.

Videotapes are being used in some experimental research to validate the results of paper and pencil tests. A student of the effectiveness of television commercials has run some preliminary tests of a method which required the cooperation of friends. He called friends and asked them to send their secretaries to his office on the pretext of picking up a package. After arriving, the secretary was asked to wait in a reception room which contained newspapers, magazines and a turned on television set. A hidden television camera monitored her behavior as she turned to the printed material watched television or just sat. She leaves, unsuspecting, and is subsequently interviewed by standard questionnaire methods to determine her exposure to television commercials and magazines and newspapers. This is a more advanced approach of the dangerous one-way mirror setting, and provides a medium for a good check on observation and self-report data.

For some reason, still photography never had much of a vogue as research hardware. Boring (1953) mentions that "Voliva supported his theory of the flat earth by a photograph of twelve miles of the shore line of Lake Winnebago: you could see, he argued, that the shore is horizontal and bowed."[1] Perhaps the obvious content and population restrictions of still photography scared investigators off.

Movie film is a different story. Earlier, we mentioned the work of Hess and Polt (1960) and Foote (1962), who used movie cameras to demonstrate pupil dilation as a physical indicator of interest. The physical location of eyes was used by Politz (1959) in a study of commercial exposure of advertising posters placed on the outside of buses. Politz' equipment were movie cameras placed in buses, which were automatically activated in a series of short bursts spread throughout a day. The camera was faced outward, over the poster under study, and, in a switch on the Bunker Hill advice, a person whose both eyes could be seen in the developed prints was classified as "in" the advertising audience for the poster. This design, which used a random sample of both locations and time, is exemplary for its control of a large number of extraneous variables that could jeopardize external validity. A question is the visibility of the camera. It occupied a bus seat, and there is the possibility that it attracted some attention. That is, the eyes counted as looking at the poster were looking at the camera instead. This should result in an overestimate of the size of the bus poster advertising audience. If one can assume that the novelty of the camera will wear off over time, one could test the hypothesis by making a longitudinal analysis of the material. If the test materials were controlled for novelty themselves, the estimated audience level should decline over the time period if a significant number of people were viewing the camera and not the poster.

---

[1] In the same article, Professor Boring comments on changes in the human observer as an instrument. "I remember how a professor of genetics many years ago showed me published drawings of cell nuclei dated both before and after the discovery and description of chromosomes. Chromosomes kept showing up in the later drawings, not in the earlier."

Maccoby (1963) also used eye movement in some work on nonverbal communication in the classroom. She filmed students and teachers simultaneously, and the initial analysis showed three variables to be important in this communication process. First, the fixity of the student's orientation toward the teacher--the focus of his eyes and the attention of his body. High fixity of postural attitude seems to be a necessary correlate of attention and understanding. A second variable involves the movement of forehead muscles. A third general category includes facial and head gestrues such as modding the head, "lighting up the eyes," and looks of comprehension, puzzlement and so on. The difficulty with Maccoby's work is that the material was obtained by placing the cameras in full view. The cameras were not in operation all the time, and then because the cameras had silent mechanisms, the students had no way of knowing at what times they were operating. For this reason, Maccoby holds that a minimum of reactivity was involved in the procedure, although other observes might not come to the same conclusion.

Landis' and Hunt's (1939) method of studying movement responses could easily be applied in nonlaboratory settings. Shooting off a gun, the authors filmed the subject's gestural response pattern, which included such movements as the drawing forward of the shoulders, contraction of the abdomen and bending of the knees. Facial patterns included closing of the eyes and widening of the mouth. It will be remembered that Krout had found that a gesture of placing the hand to nose was correlated with fear. With the stimulus of an unexpected gunshot, the immediate responses may be independent of any contaminants due to the experimental setting.

Hardware: Physical Supplanting of the Observer

The hardware measures so far mentioned have been mainly concerned with reducing the risk associated with the human observer's fallability as a measuring instrument--his selective perceptions and his lack of capacity to note all elements in a complex set of behaviors. Another use to which hardware has been put is to obtain research entree into situations which are excluded by the usual simple observational method. Some of these content areas have been unattainable because of the privacy of the behavior, others because of the prohibitive costs of maintaining observational scrutiny over a substantial enough sample of time. Sitting in for the observer, hardware can help to resolve both problems.

"Blind bugging" via audiotapes is one such approach, and a controversial one when applied in certain settings. Jury deliberations are not observable because of standard legal restraints, but Strodtbeck and his colleagues (Strodtbeck and James, 1955; Strodtbeck and Mann, 1956; Strodtbeck, James and Hawkins, 1957) received the approval of the court and counsel from both sides to place hidden microphones in the jury room. The "uncovering" of this practice caused several states to pass laws specifically barring such recordings. The use of concealed recording devices presents ethical questions that have been underlined by Amrine and Sanford (1956) and Burchard (1957).[1]

Many pieces of hardware have been developed for measuring the level of physical activity, a variable that has been viewed as symptomatic of many things. Perhaps the earliest mention of this type of measure was made by Galton (1884) who was at that time interested in the physical equivalents of metaphorical language. He took as his example the "inclination of one person toward another." This situation is clearly seen when two such people are sitting to each other at a dinner table, according to Galton. To empirically demonstrate this in quantitative terms (Galton, 1884; Watson, 1959), he suggested arranging a pressure gauge with an index and dial to indicate changes in stress on the legs of the chair on the side nearest to the other person. Galton suggested three necessary conditions for this type of experiment: the apparatus must be effective, it must not attract notice, and it must be capable of being applied to ordinary furniture. All three of these criteria are appropriate for the more recent apparatus studies.

It is obvious that such a device may be a substitute for human observers when their presence might contaminate the situation, and where no convenient hidden observation site is available. Indeed, many of the studies discussed earlier as "simple observation" are amenable to mechanization, provided that Galton's criteria could be met.

There is F. Scott Fitzgerald's fictional account of "The Last Tycoon," who evaluated the quality of rushes (preliminary, unedited film "takes") by observing how much they made him wiggle in his chair. The more the wiggles, the poorer the movie scenes. Simple observational measures could be made by a concealed observer of twistings, but they would be clearly inferior to a more mechanical device.

---

[1] We may add as an aside that the most astute device we have heard of for concealed recording is a microphone rigged in a mock hearing aid. It works extremely well in inclining the subject to lean over and shout directly into the recording apparatus. The presence of a dangling cord does not inhibit response.

Just such a device has been employed by Kretsinger (1952; 1959).[1] He used what he terms an electromagnetic movement meter to study gross bodily movement in theater audiences--a very difficult observational setting because of inadequate illumination. Kretsinger claims that this method was "objective, essentially linear, and completely removed from the subject's awareness." The technique

> ". . . was based upon a capacity operated electronic system often used in burglar alarm applications. As modified by the author, it consisted of an oscillator detector, a D.C. amplifier, and an Esterline-Angus ink writing recorder. A concealed copper screen was located near the head of the S watching the film. As the S moved, the effective capacity of the oscillator circuit varied, changing the frequency of its oscillation. This frequency shift was converted to a change in D. C. Voltage amplified sufficiently to drive an ink pen on a moving paper chart . . . completely removed from the S's awareness" (Kretsinger, 1959, 74).

The importance of heeding population characteristics is borne out by his conclusion, "There was some evidence that the presence of girls had a disquieting effect upon the boys . . ." (p.77).

Cox and Marley (1959) devised another movement measure in their study of the restlessness of patents as a partial measure of the effect of various drugs. Their rather complex apparatus consists basically of a series of pulleys and springs, set under the springs of the bed, which record the displacement of the mattress. When the patient is perfectly motionless, the relay system does not operate, but the slightest movement will be recorded. A more simple device is possible with baby cribs, wherein the activity level of the child is measured by shaving down one of the four legs and attaching a meter which records the frequency of jiggling. This is much less fine a measure than that of Cox and Marley, for the child could move without activating the meter. For studies which don't require such fine calibration however, the simplicity of the device is appealing.

---

[1] Also see Lyle (1953).

As beds, cribs and chairs can be wired, so too can decks. Foshee (1958) worked on the hypothesis that a greater general drive state would manifest itself in greater activity. Here is a good theoretical proposition testable on a device appropriate for natural settings away from the laboratory. To measure activity, Foshee used a schoolroom desk which was supported at each corner by rubber stoppers. Attached to the platform which supported the desk was a mechanical lever arrangement which amplified the longitudinal movements of the platform. Through an elaborate transmission system, the amplitude and frequency of the subject's movement could be measured. Foshee does not mention whether the subjects (in this case a group of mental defectives) were aware of the apparatus or not, but it would seem likely that the device could be constructed to evade detection.

Dr. Jerome Schulman has produced an ingenious suggestion for measuring activity level. He proposes that subjects be given self-winding watches, all carefully controlled at the time of distribution for spring tension. When the watches are gathered back again, the spring tension is again measured and a subtractive score produced for each individual. It would be hard to conceive of how this giving and taking could be conducted in a completely nonreactive format, but as one measure imbedded in a series, it has attractions. The main question is to what degree the measure really reflects total movement. If one's hypothesis or comparison is limited to arm movement, then clearly there is no issue. But if the arm movement (indexed by the watch spring) is viewed as a representative sample of total body activity, the adequacy of the sampled behavior is under question. There may be substantial response set differences among individuals in the degree to which they swing arms, and this must be carefully controlled. There is also the question of whether or not these response sets are systematically linked to social groups. In a cross-cultural study, one would predict significant differences between Italians and Swedes on the degree to which arm movement accounts for a share of total body movement. Is there a similar difference among social class groups?

To reach into the difficult setting of a darkened movie house for the study of expressive movement, several investigators have employed infrared photography (Siersted and Hansen, 1951; Bloch, 1952; Gabriele, 1956). This type of filming seems to eliminate almost entirely the element of self-awareness of the observational apparatus. It is clearly superior to unaided observation because of the advantage of working in the dark; the brighter the light in which to see the subject, the brighter the light for them to see you. This is illustrated in Leroy-Boussion's (1954) study of emotional expressions of children during a comedy film. Although Leroy-Boussion claimed to be only a projection aide during the film, she did have to eliminate certain subjects who "seemed to be aware" of her presence as an observer. It would thus seem likely that there were other subjects aware who did not make it visible to the investigator. Putting aside the question of reducing the sample size, the more important issue is whether those who were aware (and discarded as subjects) were a selective group (the more suspicious or paranoid, for example). Infrared photography drastically reduces such selective loss. Further, it is possible to match the infrared camera with the regular projection machine so that in subsequent analysis the photographs of audience reaction can be matched easily to the specific film sequence.

The danger of relying solely on interview or questionnaire self reports is sharply illustrated in the Siersted and Hansen (1951) study. They supplemented their filming by interviewing the children who had viewed the film. There were marked differences between these responses and the filmed reactions (and also differences from hidden tape recordings of verbal comments made during the film).[1]

The estimation of attendance at an event or an institution can be carried out by planting observers who count heads. Another route is to mechanize and and count circuit breaks. The "electric eye," particularly when supplemented by a time recorder, provides a useful record of frequency of attendance and its pacing. The photoelectric cells are typically set up on either side of a doorway so that any break of the current will register a mark on an attached recording device. As Trueswell (1963) shrewdly pointed out, however, this apparatus is not free from mechanical or reactivity contaminants. Particularly when the device is first installed, it is common for people to step back and forth through the light or to wave arms and legs, thus registering three or four marks for a single entry. Another difficulty is the placement of the cells. If they are set too low, it is possible for each leg to register a separate mark as the person walks through. If the doorway is wide enough to admit two people, a couple may walk together and register only one mark. It is important to note that there are not random errors which balance out, but are constant for the method. Because of this, it is equally important that human observers be there with the mechanical device, particularly in its early period of installation, to study whether any such errors are admitted to the data.

---

[1] For some reason, the French have been leaders in research on movie hardware. Toulouse and Mourgue (1948) worked with respiratory reactions to index reactions to films, and it has even been suggested that the temperature of the room in which a film is viewed might be monitored as an indicator.

Ellis and Fryer (1959) have demonstrated the complexities possible with photoelectric cells in their study of the movements of children with severe neuropathology. Their apparatus consisted of a square plywood enclosure in which the children played. Electronic devices were located on the outside surface of the walls and arranged so that the beams crisscrossed the enclosure at two-foot intervals. Interruption of a beam would be recorded, with each beam being recorded separately.

With the light beams visible, the behavior of subjects may be modified--either because they dart back and forth in a playful game with the beam, or because it inhibits their movement to know that they are under observation. Ellis and Fryer suggest modifications of their technique to avoid such risks, among them installing infrared exciter lamps, noiseless relays and soundproofing.

As a final example of the use of hardware to get otherwise difficult content, there is Weir's (1963) report of her audiotape recordings of a two and a half year old boy falling to sleep. The child practices language, working with noun substitution and articulation. In the evening of the day when he was first offered raspberries, he says, "berries, *not* bayries, berries."

Maccoby (1964) who summarized the study, states: "These observations provide insight into language learning processes which are ordinarily covert and not accessible to observation" (p. 211).

The Intervening Observer

The observation studies reported so far have been dominantly ones in which the observer is passive. He may take the behavior as it comes, or he may introduce mechanization to improve the accuracy and span of his observations, but he has not typically altered the cues in the environment to which the person or group is responding. This passivity has two costs. It is possible that the behavior under study occurs so infrequently that an inordinate amount of effort is expended on gathering large masses of data, only a small segment of which is useful. Or, paying the second cost, the naturally occurring behavior is not stimulated by events of sufficient discriminability. The investigator may want four or five levels of intensity of a condition, say, when the convenient simple observation approach can produce only two.

Rather than pay these costs, many investigators have actively stepped into the research environment and "forced" the data in a way that did not attract attention to the method. In some cases, this has meant grading experimental conditions over equivalent groups, with each group getting a different "natural" treatment. In a smaller number of studies, the conditions have been varied over the individual. Both classes are illustrated in this section.

Perhaps the most visible of the hidden observers, Allen Funt of the television show Candid Camera, gave up simple observation because of the high dross rate (Flagler, 1960). In the early years of the program, Funt's episodes consisted largely of studies of gestures and conversation (Hamburger, 1950; Martin, 1961). Particularly with conversation, Funt found that a large amount of time was required to obtain a small amount of material, and he turned to introducing confederates who would behave in such a way to direct attention to the topic of study. In one magnificent sequence of film, Funt has prepared a cross-cultural comparison of how men from different countries respond to the request of a female confederate to carry a suitcase to the corner. Filmed in the different countries, the episodes center on the girl indicating she has carried the suitcase for a long while and would like a hand. The critical material is the facial expression and bodily gestures of the men as they attempt to lift the suitcase and sag under the weight. It is filled with metal. Funt has offered to open his extensive film library to social scientists. For students of response to frustration or unexpected stimuli, this is rich ground.

Obviously, manipulation by itself is not a contaminant. It is only when the manipulation is seen as such that reactivity enters to threaten validity. Carroll (1962) showed that active initiation of stimuli can have its comic side. In an exploratory venture, he sent out twelve wires to distant friends, congratulating each on his "recent achievement." Back came eleven acknowledgments of humble thanks. This approach lacks control, for we cannot know how many acknowledged in puzzled courtesy, and how many felt Carroll had done them their due. If the study is replicated, it might be well to send a control sample a wire saying, "It doesn't matter. We're with you anyway." The wire was an efficient way to stimulate a response. An analogue may be an attempt to teach automobile driving by operant conditioning procedures. It is possible, but may take a hazardously long time contrasted to active control by the teacher (Bandura, 1962).

Simple observation, mechanized or not, is appropriate for a broad range of imaginative and useful research comparisons. Some of these we have mentioned. The advantage of contrived observation is to extend the base of simple observation and permit more subtle comparisons of the intensity of effect.

The early work of Hartshore and May yields good examples of the manipulating observer. In Studies in Service and Self Control (Hartshorne and May, 1929), they report on a long series of experiments--the first behavioral studies of "service." Employing "production methods of measurement," they indexed service or helpfulness by the subject's willingness to produce something--a toy in a shop, or the posting of a picture. Similarly, they employed "self sacrifice" techniques, measures on which the subject had to give up something.

The subjects were school children and the active involvement of the experimenter (teacher) in defining alternatives of behavior was both expected and normal. The awareness of being tested threat is less an issue in educational research, and the long line on lecture vs. discussion methods and the current research on educational television are a fine source of learning research because the risk of the contaminant is so reduced.

In the same way, it is not a patently false condition for a teacher to present students with the chance to help some other children in hospitals. Hartshore and May graded the opportunities to help in an "envelope test." The student could put pictures, jokes or stories in envelopes to give to hospitalized children, could promise to do so, or not do so. In another behavioral measure of sacrifice, one with more artificiality however, the students were told they would be given some money. They were provided opportunities to bank it, give it to a charity, or keep it themselves. In another phase, one that presaged many small group experiments, the children were given a choice as to whether they would work for themselves or for the class in a spelling contest.

Contrived observation, then, is observation in which the stimuli and/or the available responses are varied in an inconspicuous way. For Hartshorne and May, the variation was primarily of the response alternatives.

The recent series of studies by Fantz and his colleagues (Fantz, 1961a; 1961b; Fantz et al., 1962; Fantz, 1963) shows the more usual variation of the stimuli. They, too, worked with subjects where the reactivity risk is low--newborn infants--in research on pattern vision. The simple response measure was visual fix on a target, with each child exposed to six color and pattern variations. As far as a 48 hour old infant is concerned, a series of concentric rings in his visual field is as natural as anything else.

Stechler (1964) also studied newborns, observing the effect of medication administered to the mother during labor on the baby's attentiveness. Each child received three stimuli, "held near the baby's face for a total of nine minutes . . . . An observer hidden from the babies recorded the total time they looked at and away from the stimuli" (p. 315).

A much more hardheaded group of subjects, automobile salesmen, were studied by Jung (1959; 1960; 1961; 1962; 1963) in his evaluations of the effect of various bargaining postures. The response measure was simple, the quoted price of an automobile with specified features, and three different bargaining postures were struck. In this well designed series of experiments, confederates posed as customers and adopted one of these three poses: an eager, naive "I just got my license. Where do I sign?" approach; An engineering and price sophisticated approach and one between. The differences are statistically significant but smaller in absolute dollars than might have been expected--only $200 between the extremes in one study. Because the studies were conducted in the heavily price competitive Chicago area, the dollar differences may well be less than in areas in which competition is along other lines. The Chicago buyer, real or feigned, gets an automatic discount without any haggling.

Franzen (1950) conducted a similar experiment with pharmacists. The confederate was again a "customer," who related various symptoms of illness to the pharmacist. The symptoms were graded by their severity as told by the "customer," and note was taken of whether or not a visit to a physician was suggested. Examples of the symptoms, all of which could be related to early cancer, are:

loss of voice
sore on lip
heartburn and stomach trouble
constipation

Franzen also administered opinion questionnaires to an equivalent, randomly selected group of pharmacists. The results from the questionnaire are different from those of the field study. The contrived observation results, we suspect, have a higher predictive value than those from the questionnaires.

These findings recall the classic study comparing verbal attitudes and overt acts: LaPiere's (1934) research on prejudice. He, his wife and a Chinese student visited 250 hotels and restaurants and were refused service just once. Yet when questionnaires subsequently were sent to those same places, asking if Chinese customers were welcome, some 92 per cent answered negatively. As a control, LaPiere sent identical questionnaires to 100 similar establishments which his party had not visited and the response was similar.

The violation of prohibitions has offered the subject matter for several field studies of contrived observation. In these, the physical world of the observed was actively manipulated, and the dependent variables were simple motor acts.

Freed, Chandler, Mouton and Blake's (1955) experimented with sign violation. To what extent will students violate a sign urging the use of an inconvenient side door rather than the customary main door of a university building? The degree of prohibition was varied on the sign (high, medium or low) with the interesting fillip of a confederate who conformed to the sign or violated it. In a control condition no confederate was present. Ninety subjects were assigned to nine different experimental groups, combinations of the two conditions of prohibition of sign and confederate's presence or behavior. They showed main effects for the two conditions with no interaction between them.

Violation of traffic signals was the dependent variable in a study by Lefkowitz, Blake and Mouton (1955). An experimenter was again an active element in the setting--a male who dressed in either high or low status clothing. The confederate either conformed to or violated a traffic signal that ordered him to "wait" on a street corner. An observer 100 feet from the corner noted the number of people on the corner who went along with the confederate. Would the difference in dress elicit differences in the number of pedestrians violating the light? Austin, Texas was the locale and pedestrians violated the sign more often in the presence of a model, and did so significantly more often when the nonconforming model was in his higher status dress.

Cratty (1962) attempted a replication of this study in Evanston, Illinois. He added observation of the race of violaters and conformers, and his preliminary analysis suggests a significant difference between Negroes and whites on violations under both conditions. The racial composition of the sample could thus confound comparisons.[1] Cratty's findings also illustrate the problem of the cross-sectional stability of a research measure. To have good intersectional comparisons, the degree of usual conformity to the signs should be equivalent. Comparison of the Texas and Illinois data is muddied by a law-abiding difference. Only 1 per cent of the Texas sample violated the sign when the confederate was absent; over 60 per cent of the Illinois sample did.

---

[1] It might be interesting to draw a sample of U.S. cities with varying degrees of reported racial tension and compare, by race, the extent to which minor violations such as walking on a red signal are committed.

Moore and Callahan (1943) conducted research in New Haven, concentrating on traffic and parking behavior. Three sets of studies are reported, all of them nonreactive, all examples of contrived observation. In the first set, the "parking" studies, observations were made over a four-year period. They first observed the number and length of time parked, of car in areas where there was no formal ordinance against parking. Then a formal ordinance prohibiting parking was introduced and the observation continued. Thus, there is a time-series analysis in which the baseline is the period precedent to the introduction of signs.

In a second set of "administrative" studies, the militancy of enforcement of traffic violations was the independent variable. Parking was limited by ordinance (sign) to 30 minutes. In one condition, violations were tagged after 80 minutes and in a second condition, tagged after 45 minutes. The frequency and duration of parkings under these two conditions were compared to a control condition in which no tickets were given at all.

The third set centered around rotary traffic. All studies in this group used a large circular tract of pavement from which five streets radiated. Observation consisted of noting the path in which the area was crossed.

Five observation periods were employed. The first was before there was any ordinance regulating the direction of traffic. The second was before a formal ordinance was enacted, but in the presence of signs to "keep right."

7

The third followed enactment of the ordinance. The fourth was while the ordinance was visibly in effect--barriers would not allow cars through the center of the circle. Finally, the barrier and signs were removed when the ordinance was still in effect. The observed direction of flow under the five conditions was compared.

Traffic behavior offers a splendid opportunity for naturalistic experimentation. Much control data is already available, produced for engineering studies or by market research firms to document the exposure of outdoor advertising. It should be possible, for example, to study the effect of different degrees of threat in a persuasive message by studying the degree to which drivers slow down, if they do, after passing different classes of signs--say those threatening legal enforcement or the danger to personal safety. Radar devices or filming from an overhead helicopter could provide the speed measure. What may be of particular interest in such a study would be the extinction rate under the different conditions. Does an enforcement warning have a faster rate of decay than a personal safety message? What happens if both classes are equated for initial effect on speed?

Another study relating to legal processes is that of Schwartz and Skolnick (1962). They investigated the effect of criminal records on the employment opportunities of unskilled workers. Four employment folders were prepared for an applicant; all folders were identical save for description of the criminal court record of the applicant. Each of 100 employers was assigned one of four treatment folders and the employer was asked whether he could "use" the man described in the folder. The employers were never given any indication that they were participating in an experiment. Whether the applicant was described as having no record, having been acquitted with an excusing letter from a judge, acquitted without a letter or simply described as convicted had a significant effect in the incidence of employers who thought they might use the applicant.

Entrapment

Hartshorne and May's (1928) study of character, employed a measure which gave subjects a chance to cheat. (See also May and Hartshorne, 1925.) This entrapment task has been followed by others, recently by Freeman and Ataov (1960). They contrived a situation in which *S*s had a chance to cheat by grading their own examinations by a scoring sheet. Using three classes of questions--fill in blanks, multiple choice and true-false--they found the number of changed answers for the three classes of questions. The three formed a Guttman scale with a reproductibility coefficient of .94.

Grinder (1961; 1962) and Grinder and McMichael (1963) have reported studies using a "ray gun" type of apparatus. Like Hartshorne and May, their interest was in studying character, or "conscience development." Children operated a "ray gun" shooting gallery individually in a realistic game situation.

"Seated seven feet from a target box, subjects were asked to shoot the ray gun pistol 20 times at a rotating rocket. With each pull of the ray gun trigger, prearranged scores from zero to five were registered by score lights also housed in the target box. High scores were rewarded with a marksman, sharpshooter, or expert badge . . . *S*s cumulated their scores on a paper score sheet. Subjects were judged to have resisted temptation if the scores recorded on their score sheets indicated that they had not earned a badge (they could not honestly). They were judged to have yielded to temptation if their score sheet showed that they had falsified their scores in order to earn one of the badges" (p. 504).

Similar in its assumption of dishonesty by subjects is a study by Brock and Guidice (1963). The subjects were students ranging from the second to the sixth grade, and these children were individually asked if they would leave the class, go to another room and partidpate in an experiment. Upon entering the room, the subject found the experimenter in a flustered state with her purse spilled and money lying on the floor. At this point the experimenter left, saying she would be back shortly, and asking the subject to pick up the contents of her purse while she was gone. The measure used was the amount of money stolen by the subject. Again, as in the Grinder and McMichael (1963) study, reactive measures were used as well to correlate the thievery measure with scales of temporal conception. There is a good risk that some of the children, having been told to go to another room for an "experiment," became suspicious. The effect of this is to make uncertain how many of those who took no money were honest and how many were acute. This bias could possibly interact with a practice effect on the part of the experimenter. Did she play her role differently over time, as she became more practised or more bored? If professional actresses have problems "keeping fresh" it is reasonable to ask about amateurs.

The more complex the action of the confederate, the greater the risk of an experimenter effect, and the greater the possibility for more gradations in the experimental variable.

Petition signing has been the dependent variable in observational studies by Blake, Mouton and Hain (1956) and Helson, Blake and Mouton (1958). In the the first study, the strength of the plea to sign was varied, and frequency of signing was closely associated with this variable. A confederate was then introduced into this situation to provide varying reactions of another person. In some situations he signed readily, in others refused and in still others his response was unknown to the person approached. The behavior of the confederate influenced signing, and an interesting finding was that both variables operated independently--the strength of the plea and endorsement by another.

The Helson, Blake and Mouton (1958) study used petition signing as a response behavior within a larger experimental setting. Students who had volunteered to participate in an experiment were taken by a guide to an experimental room. On the way, the pair was stopped and a confederate asked the guide to sign a petition. After the confederate signed or did not sign, the subject was asked to sign. Four conditions were distributed, with each student receiving one:

1. A petition on a proposal that had previously elicited 96% positive response and confederate signed.
2. The same 96% positive proposal, but confederate refused to sign.
3. A proposal which had received only a 15% positive response and confederate signs.
4. Fifteen per cent proposal; confederate doesn't sign.

Solicitation to sign a petition is a common enough event in academic settings, and may be becoming more common outside the cloistered world. Certainly it offers a broad freedom of movement in experimentation and structuring of contrived conditions. The searching for volunteers does too, and several studies have used observation of the simple "volunteer-not volunteer" as the behavioral variable.

Schacter and Hall (1952) experimented with college students, employing different situational elements in eliciting volunteers for an experiment and then noting whether or not the volunteers did in fact show up for the experiment. Classes were divided into four groups and given different restraints. In one group, after a requesting speech, the listeners were asked to fill out a questionnaire, whether or not they wished to participate in the experiment. Those who did wish to participate were merely asked to check an appropriate place on the form. In a second group, the forms were passed around the room, and anyone who wanted to participate in the experiment could take one. A third group was asked to raise hands if interested and a fourth group also raised hands, but half of the class had been enlisted as confederates and "volunteered." Schacter and Hall's conclusion was that neither the high nor low constraint condition was particularly desirable in soliciting volunteers. If they made it easy to refuse, they got a high refusal rate, but high attendance among those who did volunteer. Contrariwise, placing high pressure on volunteering yields a higher level of volunteers, but a smaller number who actually perform on their promise.

Rosenbaum and Blake (1955) wanted to test the hypothesis that the act of volunteering is "a special case of conformance with social norms or standards, rather than . . . an individualistic act conditioned by an essentially unidentifiable complex of inner tensions, needs, etc." (p. 193). Subjects were plucked from students studying in the university library and conditions were varied so that the subject either saw an acceptance or rejection of the request from a confederate. In a third group, the student accepted or rejected the volunteering request in the absence of model. As predicted, acceptances were high with a conforming confederate, low with a nonconforming confederate and in between on the control condition.

The University of Texas library was also the site of another study by Rosenbaum (1956). Volunteering was the dependent variable and stimuli conditions included three request strengths (determined by a pilot study) and three background conditions employing confederates. The confederate entered the library, sat next to an unsuspecting student and then the volunteer seeker entered and started with the confederate.

Blake, Berkowitz, Bellamy and Mouton (1956)determined the effect on the level of volunteering by varying the attractiveness of alternatives to volunteering. The public or private character of the volunteering was also varied, with conditions altered so that a class might substitute volunteering time for time otherwise devoted to: (1) a pop quiz, (2) released time from class, (3) a control, no time gained. It might be observed that under the pop quiz alternative, 98.8% of Ss volunteered under the private commitment situation and 100.0% under public commitment. That 1.2% above is accounted for by the single aberrant student who preferred a pop quiz to participation in an experiment.

Volunteering for social action was the subject of a study by Gore and Rotter (1963). The action in this case was the willingness of students in a Southern Negro college to engage in different types of segregation protest activity. This criterion measure was correlated with previously obtained scores on control of reinforcement and social desirability scales, the latter scales not specifically dealing with the segregation issue. A generalized attitude toward internal or external control was shown to be predictive of the type and degree of commitment behavior performed in an attempt to effect social change.

An Over-All Appraisal on Hidden Hardware and Control

As the discussion and examples of observational research had progressed over these past two chapters, the reader may have been sensitive to a movement along a passivity-activity dimension. With the studies reported in the chapter on simple observation methods, the observer was a nonintervening, passive onlooker of behavior that came before his eyes or ears. He may have scrambled about in different locations to reduce some population restrictions, but his role was a more quiet, receptive one. In many ways this is appropriate for the covert character of the studies we have outlined.

With the hardware employed in the studies cited in this chapter, the investigator engaged his data more--actively expanding the possible scope of the content of research and achieving a more faithful record of what behaviors did go on. Yet the hardware varied too. Some of the hardware devices are static while others are mobile. To the degree that the hardware is mobile (say a microphone in a mock hearing aid _v._ one secreted in a table), the experimenter has flexibility to make more economical forays into locational sampling. He could sample in a number of locations by installing more permanent recording devices, but commonly a more feasible method is to sample occasions and time with mobile equipment. As electronic technology develops, more opportunities arise. It isn't so long ago that television cameras had to be anchored in one spot.

When, through deliberate choice or no realistic alternative, the investigator was limited to a fixed instrument, he was forced to depend on the character of the population which flowed past that spot and the content which was appropriate to it. The waiting game can give accurate and complete measurement of a limited population and limited content, and the decision to use such an approach is posited on two criteria, one "theoretical," one "practical." Are the limitations likely to be selective enough to inhibit the generalizability of the findings? Can the investigator absorb the time and money costs of developing material with a low saturation of pertinent data for his comparisons?

In the contrived observation studies, the experimenters took the next step and actively intervened in the production of the data, striding over the line from passive, critically placed observations. The effect of this was to produce very dense data, material of which a high proportion was pertinent to the research comparisons. Further, a finer gradation of stimuli was then possible and more subtle shadings of difference could be noted. By the active intervention, as the petition and volunteering studies of conformity show, it was also possible to make estimates of the interaction of variables, an extremely difficult matter with passive observation.

As the experimenter's activity increases, and he achieves the gains of finer measurement and control, the price he pays is the increased risk that he will get caught, that the subjects of the observation will detect the recording device, will suspect that the confederate is really a "plant." This is a high price, for if he is detected, his research is flooded with the reactive measurement errors which the hidden observation approach, regardless of its simplicity or complexity, is designed to avoid. At the extreme end of contrivance, when a confederate is a visible actor in the subject's world, it requires the greatest finesse to protect against detection and against comparison damaging changes in the behavior of the confederate. The best defense, as always, is knowledge and almost all of the observational approaches have built into them the capacity to examine whether or not population or instrument contaminants are working to confound the data. This is the not inconsiderable gain of obtaining your data first hand and seeing it develop with your own eyes.

Chapter 7

A Final Note

"In the dialectic between impulsivity and restraint, the scientific superego became too harsh--a development that was particularly effective in intimidating adventurous research, because the young were learning more about methodological pitfalls than had their elders. . . ."

(Riesman, 1959,p. 11).

David Riesman's remarks on the evolution of communications research apply equally well to the broader panoply of the study of social behavior. As social scientists, we have learned much of the labyrinthe that is research on human behavior, and in so doing discovered an abundance of cul-de-sacs. Learning of the complexities of the maze shortened our stride through it, and often led to a pattern of timid steps, frequently retraced. No more can the knowledgeable person enjoy the casual bravura that marked the sweeping and easy generalizations of an earlier day.

The facile promulgation of "truth," backed by a few observations massaged by introspection, properly met its end--flattened by a more questionning and sophisticated rigor. The blackballing of introspection was a positive advance, but an advance by subtraction. Partially as a reaction to the grandiosities of the past, partially as a result of a growing sophistication on the opportunities for error, the scope of individual research studies shrank, both in the range of content considered and the diversity of procedures.

The shrinkage was understandable and desirable, for certainly no science can develop until a base is reached from which reliable and consistent empirical findings can be produced.[1] But if reliability is the initial step of a science, validity is its necessary stride. The primary effect of improved methodological practices has been to further what we earlier called the internal validity of a comparison--the confidence that a true difference is being observed. Unfortunately, practices have not advanced so far in improving external validity--the confidence with which the findings can be generalized to populations beyond those immediately studied.

Depressing this advance in ability to generalize was the laissez-faire intellectualism of the operational definition. Operationalism definitionalism (to use a ponderously cumbersome term) provided a methodological justification for the scientist to not stray beyond a highly narrow, if reliable, base. One could follow a single method in developing data and be "pure," even if this purity were more associated with sterility than virtue.

The corkscrew convolutions of the maze of behavior were ironed, by definitional fiat, into a two dimensional T maze. To define a social attitude, for example, solely by the character of responses to a list of questionnaire items is eminently legitimate--so much so that almost everything we know about attitudes comes from such research. Almost everything we know about attitudes is also suspect because the findings are saturated with the inherent risks of self-report information. One swallow does not make a summer; nor do two "strongly agrees," one "disagree" and an "I don't know" make an attitude or social value.

Questionnaires and interviews are probably the most flexible and generally useful devices we have to gather information. Our criticism is not against them, but against the tradition which permitted them to become the methodological sanctuary to which the myopia of operational definitionalism permitted a retreat. If one were going to be limited to a single method, then certainly the verbal report from a respondent is the choice. With no other device can an investigator swing his attention into so many different areas of substantive content, often simultaneously, and also gather intelligence on the extent to which his findings are hampered by population restrictions.

---

[1] "Almost all experiments on the effects of persuasive communications, including those reported in the present volume, have been limited to investigating changes in opinion. The reason, of course, is that such changes can readily be assessed in a highly reliable way, whereas other components of verbalizable attitudes, although of considerable theoretical interest, are much more difficult to measure" (Janis & Hovland, 1959, 3).

The power of the questionnaire and interview has been enormously enhanced, as have all methods, by the development of sensitive sampling procedures. With the early impetus provided by the Census Bureau on locational sampling, particularly on the theory and practice of stratification, concern about the population restrictions of a research sample have been radically diminished. Less well developed is the random sampling of time units--either over long periods such as months, or within a shorter period such as the day. There is no theoretical reason why time sampling is scarce, for it is a simple question of substituting time for location in a sampling design. Time sampling is of interest not only for its control over population fluctuations which might confound comparisons, but also because it permits control over the possibility of variable content at different times of the day or different months of the year.

The cost is high. And for that reason, government and commercial research organizations have led in the area, while academic research continues to limp along with conscripted sophomores. The controlled setting makes for excellent internal validity, as one has high control over the conditions of administration and the internal structure of the questionnaire, but the spectre of generalizability is everpresent.

That same spectre is present, however, even if one has a national probability sample and the most carefully prepared questionnaire form or interview schedule. So long as one has only a single class of data collection, and that class is the questionnaire or interview, there is inadequate knowledge of the rival hypotheses grouped under the term, "reactive measurement effects." These potential error sources, some stemming from an individual's awareness of being tested, others from the nature of the investigator, must be accounted for by some other class of measurement than the verbal self report.

It is too much to ask of any single class that it eliminate all the rival hypotheses subsumed under the population, content and reactive effects groupings. So long as the research strategy is based on a single measurement class, some flanks will be exposed, and if fewer are exposed with the choice of the questionnaire method, there is still insufficient justification for its use as the only approach.

If no single measurement class is perfect, so too are none alien or illegitimate. Many studies and many novel sources of data have been mentioned in the earlier chapters. The reader may indeed have wondered which turn of the page would provide a commentary on some Ouija board investigation. It would have been there had we known of one, and had it met some reasonable criteria of scientific worth. These "oddball" studies have been discussed because they demonstrate ways in which the investigator may shore up reactive infirmities of the interview and questionnaire. As a group, these classes of measurement are themselves infirm, and individually contain more risk (more rival plausible hypotheses) than does a well-constructed interview.

This does not trouble us, nor does it argue against their use, for the most fertile search for validity comes from a combined series of different measures, each with its idiosyncratic weaknesses, each pointed to a single hypothesis. When a hypothesis can survive the confrontation of a series of complementary methods of testing, it contains a degree of validity unattainable for one tested within the more constricted framework of a single method. Findings from this latter approach must always be subject to the suspicion that they are method bound: will the comparison totter when exposed to an equally prudent but different testing method? There must be a multiple operationalism.

E. G. Boring (1953) put it this way:

> ". . . as long as a new construct has only the single operational definition that it received at birth, it is just a construct. When it gets two alternative operational definitions, it is beginning to be validated. When the defining operations, because of proven correlations, are many, then it becomes reified" (p. 222).

This means, obviously, that the notion of a single "critical experiment" is erroneous. *There must be a series of linked critical experiments, each testing a different outcropping of the hypothesis*. It is through triangulation of data procured from different measurement classes that the investigator can most effectively strip of plausibility rival explanations for his comparison. The usual procedural question asked is, which of the several available data collection methods will be best for my research problem? We suggest the alternative question: which *set* of methods will be best?--with "best" defined as a series which provides data to test the most significant threats to a comparison and do so with a reasonable expenditure of resources.

There are a number of research conditions which might hold in which the sole use of the interview or questionnaire leaves unanswerable rival explanations. The purpose of the less popular measurement classes discussed here is to bolster these weak spots and provide intelligence which will evaluate validity threats. The payout for using these measures is high, but the approach is more demanding of the investigator. In their discussion of statistical records, Selltiz et al. (1959) note:

> "The use of such data demands a capacity to ask many different questions related to the research problem . . . .. The guiding principle for the use of available statistics consists in keeping oneself flexible with respect to the form in which research questions are asked" (p. 318).

This flexibility of thought is required to handle the reactive measurement effects--which are the most systematic weakness of all interview and questionnaire studies. These error threats are also systematically present in all observation studies in which the observer is known and/or visible to those under study. To varying degrees, measurements conducted in natural settings, without the individual's knowledge, control this type of error possibility. In all of them, hidden observation, contrived observation, trace analysis and secondary records, the individual is not aware of being tested, and there is little danger that the act of measurement will itself serve as a force for change in behavior and/or elicit role playing that confounds the data. There is also minimal risk that biases coming from the physical appearance or other cues provided by the investigator will contaminate the results.

In the observational studies, however, hiding the observer does not eliminate the risk that he will change as a data collecting instrument over time. Any change, for the better or worse, will introduce shifts that might be erroneously interpreted as stemming from the causal variable. This source of reactivity must be guarded against in the same way that it is in other measurement classes--by careful training of the observer (interviewer) and by permitting practice effects to take place before the critical data are collected. There is no way of knowing, of course, whether all reasonable precautions have worked. For this, the only solution is an internal longitudinal analysis of data from a single observer and cross-analysis of data from different interviewers at various time points during the data collection.

Finally, none of the methods emphasized here, by themselves, can eliminate response sets which might strongly influence the character of the data. These must be brought under experimental control by manipulation of the setting itself (as in contrived field experimentation) or by statistical operations with the data if the character of the response sets are known well enough to permit adjustments. With archival records, it may be extremely difficult to know if response sets were operating at the time the data were produced.

Another weakness of the interview and questionnaire (as well as one of its strengths) is its dependence upon language. When one is working within a single society, there is always the question of whether the differential verbal skills of various subcultures will mislead the investigator. It is possible, if the groups vary in articulateness, to overgeneralize the behavior or attitudes of the group with the greater verbal fluency. This risk is particularly marked for readers, in their interpretation of a research report which liberally employs quotes. The natural tendency of the writer is to choose illustrative quotes which are more fluent, more dramatic or more engaging. If the pool of good quotes is variable across the subcultures, the reader may mistakenly overvalue the ideas in the quotes, even though the writer himself does not. This is a presentation question, but an important one because of the disproportionate weight that may be placed on population segments.

The differential capacity to use the language artfully is one source of error, while the absolute capacity of the language to convey ideas is another.[1] This is an issue strongly present in cross-cultural comparisons, where different languages may vary radically as a medium of information transfer. The effect of this is to limit the content possible for study with questionnaires or interviews. If one worked in New Guinea, for example, and had to depend upon the lingua franca pidgin widely spoken there, he would find it adequate to indicate an answer to "Where do you keep your fishing nets?" but too gross a filter to study the ethnocentricism of a tribe. Pidgin simply does not possess the subtle gradients required to yield textured responses to questions on attitudes toward neighboring tribes or one's own tribe. Although it is theoretically possible to learn all the regional dialects well enough to be competent in a qualified language medium, in practice this does not occur. A more pragmatic approach is to search for observational or trace evidence which will document aspects of ethnocentrism and then relate it to the verbal responses in the inadequate pidgin.

One more weakness of the dependence on language is that sometimes there is silence. So long as a respondent talks, glibly or not, in a rich language or not, checks and controls can be worked on the reported content.[2]

---

[1] In a similar note on observers, Heyns and Lippitt (1954) ask if the "observer lacks the sensitivity or the vocabulary which the particular observation requires."

[2] For an extended discussion of this issue, see Hyman, Cobb, Feldman, Hart and Stember (1954) and Kahn and Cannell (1957).

There are, however, situations in which refusals to cooperate preclude any chance of correcting distorted information. This usually results in a biased research population, and not a rejection of all findings, because it is almost always possible to find some people who will discuss any topic. But it can also result in a complete stalemate if only the verbal report is considered as the research instrument. An amusing example of this inability to get data by verbal report, and a nonreactive circumvention, is provided by Shadegg (1964). In his book on political campaign methods, Shadegg talks of a campaign manager who used every available method to learn the plans of his opponent--who, reasonably enough, was unwilling to grant a revealing interview. One method arranged for procuring the contents of his opponent's wastebasket: "He came into possession of carbon copies of letters . . . memos in the handwriting of his opponent's manager."[1] Admittedly a less efficient method than the interview, it met the admirable criterion of being workable: "It took a lot of digging through the trash to come up with the nuggets. But . . . daily panning produced some very fine gold." The "investigator" did not limit himself to inferences drawn from observations of his opponent's public acts, but was able to develop ingeniously (although perhaps not ethically) a trace measure to complement the observation. Each aids the other, for the observations give a validity check on the nuggets among the trash (was misleading material being planted?) and the nuggets gave a more accurate means to interpret the meaning of the public acts.

---

[1] Empirical evidence of how others are sensitive to wastebaskets as a source of inference is the practice in diplomatic embassies of burning their refuse under guard.

Generally speaking then, observational and trace methods are indicated as supplementary or primary when language may serve as a poor medium of information--either because of its differential use, its absolute capacity for transfer or when significant elements of the research population are silent.

The verbal methods are necessarily weak along another dimension, the study of past behavior or of change. For historical studies, there is no alternative but to rely mainly on records of the past time. Behavioral research on the distant past is rare however; more common are studies which center on experiences within the lifetime of respondents. There is a large literature, for example, on child rearing practices, in which mothers recollect their behavior of years past. A sole dependence on this type of data gathering is highly suspect. It may be enough to note that Thomas Jefferson, in his later years, observed that winters weren't as cold as they used to be.[1] Available records could be used to check both Mr. Jefferson and other observers of secular changes in winter's fierceness.

When, through death or refusal, reports of past behavior are unavailable, a proper contingent strategy is to interview others who have had access to the same information, or who can report second hand. This is very shaky information, but useful if other intelligence is available as a check. For many investigations, of course, the nature of the distortion is itself an important datum and can become a central topic of study when a reliable baseline is possible.[1] If other materials are present, and they usually are in a record keeping society, the best way to estimate past behavior is to combine methods of study of archival records, available traces and verbal reports, even if secondhand. Clearly, the direct observational methods are useless for past events.

---

[1] For more current evidence on the fallability of such recall data, see Pyles, Stolz and Macfarlane, 1935; McGraw and Molloy, 1941; Smith, 1958; Weiss and Dawis, 1960--all of whom comment on,or test, the validity of mothers' recall of child rearing practices. Weiss and Dawis said, "It is indefensible to assume the validity of purportedly factual data obtained by interview" (p. 384). The work of Haggard, Brekstad and Skard (1960) and Robbins (1963) suggest that it is a problem of differentially accurate recall. In Haggard's phrase, the interviews "did not reflect the mothers' earlier experiences and attitudes so much as their current picture of the past" (p. 317).

With studies of social change, the most practical method is to rely on available records, supplemented with verbal recall. If one wanted more control over the data, it would be possible to conduct a continuing series of field experiments extending over a long period of years. But the difficulty of such an approach is evidenced by the scarcity of such longitudinal, original data studies in social science. Forgetting the number of years required, there is the problem of unstable populations over time, an even greater problem in the future as the society becomes more mobile. Potential errors lie in both directions, as one moves forward or backward in time, and the more practical approach of the two is to analyze data already collected--making the everpresent assumption that such are available.

A more integrative approach for studying change is to develop two discrete time series--one is based on available records, the other freshly developed by the investigator. With this strategy, it is necessary to have an overlap period in which the relationships between the two series are established. Given knowledge of the relationship, the available records can be studied retrospectively, thereby providing more intelligence than would be possible if they existed alone. Again, there is a necessary assumption: one must be able to reject the plausibility of an interaction between time and the method. If there is any content or population fluctuation beyond chance, such a method is invalid. Diagrammatically, where O = an observation and the subscripts n = new data and a = available data:

---

[1] The courts have handled secondary information by excluding it under the "hearsay" rule (Wigmore, 1935; Morgan, 1963). Epically put, "Pouring rumored scandal into the bent ear of blabbering busybodies in a pool room or gambling house is no more disreputable than pronouncing it with clipped accents in a courtroom" (Donnelly, Goldstein and Schwartz, 1962, p. 277). The case from which this is cited is Holmes, 379 p.a. 599 (1954).

$$O_{a1}O_{a2}\ O_{a3}O_{a4}O_{a5}O_{a6}O_{a7}O_{a8}O_{a9}O_{a10}$$

$$O_{n6}O_{n7}O_{n8}O_{n9}O_{n10}$$

The final major purpose of the less reactive methods is cost reduction. Many scholars know how to conduct massive surveys which effectively control major error sources; few do so. This knowledge is an underdeveloped resource. With survey interviews costing as much as $10 apiece, the failure is understandable, however regrettable. When the interview or questionnaire is viewed as the only method, then the researcher is doomed to either frustration or a studied avoidance of thoughts on external validity. Peace of mind will come if the investigator breaks the single method mold and examines the extent to which other measurement classes can substitute for verbal reports. The price for each individual unit of data is extremely low for most of the methods we have stressed. In some cases, the dross rate is high, and it may be necessary to observe 100 cases before one meets the research specifications. Nonetheless, even under these high dross rate conditions, the cost per usable response is often lower than that of a completed interview or returned questionnaire. The lower cost makes for flexibility to expand into content and population areas otherwise precluded, and the result of this is to increase the confidence one has in generalizing findings. Just as in the case of studying social change, it may be possible to generate different data series, some based on verbal reports, others based on secondary or observational data. Providing for enough cases of the more expensive procedures to yield a broad case for linkage, the larger number of cases can be allocated to the less expensive observational or secondary methods.

In the multimethod pattern of testing, then, the primary gains coming from the less popular methods are protection against reactive measurement threats, an expansion into content areas where verbal reports are unreliable, an easier method of determining long term change, and a lower cost substitute for some standard survey practices.

For these gains, there are associated problems for each of the less popular measurement classes--indeed if they were less problematic, we would be writing an argument in favor of an increased use of the interview.

The most powerful aspect of the verbal methods--their ability to reach into all content areas--is a soft spot with the hidden observation, trace and archival analysis procedures. We have noted remarkably adept and nonobvious applications of data from these sources, but for some content areas, the most imaginative of investigators will have trouble finding pertinent material. Individually, they are simply as broad gauged.

Often missing, too, is complete knowledge of the conditions under which the data were collected, the definitions of important terms used in classification, the control or lack of it over error risks that may be salient. This is particularly disturbing when dealing with comparisons of public records from different areas or from widely different times. The variation in definitions of a "suicide" vs. "accidental death," or the differential thoroughness with which marriages enter official records are examples of this issue.

In general, for the trace and erosion and the archival records, a dominant concern is the possibility of selective deposit and selective survival of the research data. Through supporting research designed to learn of these errors, it is sometimes possible to apply corrections to what is available. At other times, the researcher must remain in ignorance and make assumptions. If he restricts himself to *only* working with such data, he remains helpless before the vagaries of the data. If he uses other measurement classes, the process of triangulating all the different data may provide a test of his assumptions and reveal the presence or extent of error. The comparison of data from the different classes can always add intelligence unavailable from comparisons of data from within the single class.

Because of the error risks and the danger of unknown biases, we have stressed the importance of careful data sampling. Wherever feasible, locational sampling should be employed, extending over regions as well as areas within a single locality. Similarly, time sampling should be considered not only as device employed within a single day or week, but applied over months and years. Through such effort, we are able to protect against both population and content restrictions, and very often produce interesting data from comparisons of results from different locations or times.[1]

The demand for time and location sampling is no greater with archival or observational data than it is with interviews or questionnaires, for sampling is a problem that transcends the class of measurement. What is attractive about the nonquestionnaire methods is that they permit, from a practical point of view, rigorous sampling. Because the unit cost of each research datum is usually lower, an investigator has more tolerance to invest his resources in sampling designs. The absolute cost of the sampling plan is relatively constant across measurement classes. If the n is held constant, the interview study is substantially more expensive than an observational one, say, and sampling costs occupy a larger share of the research budget. If the research costs are kept constant and an equivalently rigorous sampling is conducted, one may have an interview study with a small n compared to an observational study with a large n. We suggest neither to the exclusion of the other. If links can be established across measurement classes, which requires more than one method, it may be possible to have a larger proportion of the total n accounted for by the more economical devices. The visible need for rigorous sampling is blatant with such materials as we have emphasized, but is equally real, if more masked, with other approaches.

---

[1] See, for example, Caplow and McGee's (1958) discussion of variation in salaries in American universities--particularly the relationship between beginning salaries and the prestige of the institution. In a related report, the head of an employment agency reported, "The Chicago advertising man on the average makes 10 per cent more than his New York counterpart, 25 per cent more than he would make on the West Coast, and 40 per cent more than he would make in a small town or in the south" (Anonymous, 1962, 65).

Another common demand, this one not so applicable to the verbal report approaches, is that for data transformation. The need comes from the experimenter's decreased control over the production of his materials. The exception to this is the contrived field experiment, where the investigator can have full control, but the data from archives, trace sources and observations are frequently too raw to be used as is. The need is underlined because of one of the major advantages of the secondary data--its ability to produce fine time-series information. In time series, it is usually necessary to account for extraneous sources of variation such as secular or cyclical trends independent of the effect of the variable under study. Thus the "score" which is the basis of comparison is some transformed measure which is a residual of the total "score." In other studies, the absolute number of cases varies from unit time to unit time, and the only reasonable comparison score is one which is related in some way, through an average or percentage for example, to the variable baseline. The investigator may have no control over the flow of an observed population, but he can obtain a count of that flow and use this intelligence as the basis for modifying his comparison score.

The more sophisticated forms of transformation, such as the generation of index numbers combining multiple inputs, demand more information, particularly on the relative weights which should be assigned to the various inputs before they are collected together into a single summary score. This is not so awesome as it sounds, and if the investigator is sensitive to the potential usefulness of index numbers, he often finds enough data available to provide good estimates, or that sufficient information is procurable without extraordinarily high marginal costs. Insofar as these transformations demand time and labor to make the raw data more precise, they are a disadvantage relative to standard questionnaire type procedures. There are, however, as we have suggested in various points in the text, indications that index numbers and more simple transformations could be used properly in all classes of measurement. The Zeitgeist may as yet be inappropriate, but an important work will someday link the index number theory and literature to social science measurement theory and practice.

Those, then, are the gains, those the losses. There are no rewards for ingenuity as such, and the payoff comes only when ingenuity leads to new means for more valid comparisons. In the available grabbag of imperfect research methods, there is room for new uses of the old.

Max Eastman once suggested that books should start with a first chapter consisting of one sentence, a second chapter a paragraph long and so on. He even wrote one like that--The Enjoyment of Laughter. Since this has been an unconventional monograph on unconventional research procedures, it is proper that it should have an unconventional close. We reverse Eastman's advice and offer a one phrase last chapter and a one paragraph next to last chapter.

Chapter 8

A Statistician on Method

"We must use all available weapons of attack, face our problems realistically and not retreat to the land of fashionable sterility, learn to sweat over our data with an admixture of judgment and intuitive rumination, and accept the usefulness of particular data even when the level of analysis available for them is markedly below that available for other data in the empirical area" (Binder, 1964, p. 294).

Chapter 9

Cardinal Newman's Epitaph

"From symbols and shadows to the truth."

Adams, J. S. Wage inequities, productivity, and work quality. In *Psychological research on pay*. Institute of Industrial Relations. Berkeley, Calif.: Univer. of California, 1963, Reprint #220, 9-16.

Alger, C. F. Personal communication, 1962.

Alger, C. F. Interaction in the General Assembly. Unpublished paper, delivered at International Relations Program Seminar, Northwestern University, 1963.

Allport, G. W. *The use of personal documents in psychological science*, New York: Social Science Research Council, 1942.

Amrine, M., & Sanford, F. In the matter of juries, democracy, science, truth, senators, and bugs. *Amer. Psychologist*, 1956, 11, 54-60.

Amthauer, R. Ergebnisse einer Studie über krankheitsbedingte Fehlzeiten. *Psychol. Rdsch.*, 1963, 14, 1-12.

Anastasi, A. *Differential psychology*. (3rd ed.) New York: Macmillan, 1958.

Andrew, R. J. The origin and evolution of the calls and facial expressions of the primates. *Behavior*, 1963, 20, 1-109.

Angell, R. C. The moral integration of American cities. *Amer. J. Sociol.*, 1951, 57, 123-126.

Anonymoi. Methodological inclination and hair length: an objective index of soft headedness. Unpublished manuscript, Northwestern Univer., 1964.

Anonymous. *Advertising Age*, 1962, 33, 65.

Anonymous. Senator Salinger? *Newsweek*, August 10, 1964, p. 28.

Anonymous. Z-Frank stresses radio to build big Chevy dealership. Advertising Age, 1962, 33, 83.

Argyle, M., Garner, G., & Cioffi, F. Supervisory methods related to productivity, absenteeism and labor turnover. Hum. Relat., 1958, 11, 23-40.

Arrington, R. Time sampling in studies of social behavior: a critical review of techniques and results with research suggestions. Psychol. Bull., 1943, 40, 81-124.

Arsenian, J. M. Young children in an insecure situation. J. abnorm. soc. Psychol., 1943, 38, 225-249.

Ashley, J. W. Stock prices and changes in earnings and dividends: some empirical results. J. Pol. Economy, 1962, 70, 82-85.

Asius, A. Positive and negative. U.S. Camera, 1962, 25, 32.

Athey, K. R., Coleman, J. E., Reitman, A. P., & Tang, J. Two experiments showing the effect of the interviewer's racial background on responses to questionnaires concerning racial issues. J. appl. Psychol., 1960, 44, 244-246.

Babchuk, N., & Bates, A. P. Professor or producer: the two faces of academic man. Soc. Forces, 1962, 40, 341-348.

Back, K. W. The well-informed informant. In R. N. Adams, & J.J. Preiss (Eds.), Human organization research. Homewood, Ill.: Dorsey Press, 1960.

Bain, H. M., & Hecock, D. S. Ballot position and voter's choice: the arrangement of names on the ballot and its effect on the voter. Detroit: Wayne State Univer. Press, 1957.

Bain, R.K. The researcher's role: a case study. In R. N. Adams, & J. J. Preiss (Eds.), Human organization research. Homewood, Ill.: Dorsey Press, 1960.

Bales, R.F. Interaction process analysis. Cambridge: Addison-Wesley, 1950.

Bandura, A., & Walters, R. Social Learning and Personality Development. New York: Holt, Rinehart & Winston, 1963.

Barch, A. M., Trumbo, D., & Nangle, J. Social setting and conformity to a legal requirement. J. abnorm. soc. Psychol., 1957, 55, 396-398.

Baring-Gould, W. Sherlock Holmes of Baker Street. New York: C. N. Potter, 1962.

Barker, R. G., & Wright, H. F. One boy's day: a specimen record of behavior. New York: Harper, 1951.

Barker, R. G., & Wright, H. F. Midwest and its children: the psychological ecology of an American town. Evanston, Ill.: Row, Peterson, 1954.

Barry, H. Relationships between child training and the pictorial arts. J. abnorm. soc. Psychol., 1957, 54, 380-383.

Barzun, J. The delights of detection. New York: Criterion Books, 1961.

Bass, B.M. Ultimate criteria of organizational worth. Personnel Psychol., 1952, 5, 157-173.

Baxter, J. Chicago shops pay better than N.Y.: big agencies pay more too. Advertising Age, May 28, 1962, 33, 65.

Becker, H. S. Problems of inference and proof in participant observation. Amer. sociol. Rev., 1958, 23, 652-660.

Belknap, G. M. A method for analyzing legislative behavior. Midwest J. Pol. Sci., 1958, 2, 377-402.

Beloff, J., & Beloff, H. The influence of valence on distance judgments of human faces. J. abnorm. soc. Psychol., 1961, 62, 720-722.

Benney, M., Riesman, D., & Star, A. Age and sex in the interview. Amer. J. Sociol., 1956, 62, 143-152.

Berkson, G., & Fitz-Gerald, F. L. Eye fixation aspect of attention to visual stimuli in infant chimpanzees. Science, 1963, 139, 586-587.

Berlyne, D. E. Emotional aspects of learning. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.), Annu. Rev. Psychol., 1964, 15, 115-142.

Bernberg, R. E. Socio-psychological factors in industrial morale: I. The prediction of specific indicators. J. soc. Psychol., 1952, 36, 73-82.

Bernstein, E. M. Money and the economic system. Chapel Hill, N.C.: Univer. North Carolina Press, 1935.

Berreman, J. V. M. Factors affecting the sale of modern books of fiction: a study of social psychology. Unpublished doctoral dissertation, Stanford Univer., 1940.

Binder, A. Statistical theory. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.), Annu. Rev. Psychol., 1964, 15, 277-310.

Birdwhistell, R. L. Paralanguage: 25 years after Sapir. Paper presented at Pittsburgh Bi-centennial, 1959, Research in Psychiatry.

Birdwhistell, R. L. Kinesics analysis in the investigation of the emotions. Paper presented at meeting of the American Association for the Advancement of Science, 1960.

Birdwhistell, R. L. Communicating on purpose. Keynote address: VPGA, 1963.

Blake, R. R., Berkowitz, H., Bellamy, R. Q., & Mouton, J.S. Volunteering as an avoidance act. J. abnorm. soc. Psychol., 1956, 53, 154-156.

Blake, R. R., Mouton, J. S., & Hain, J. D. Social forces in petition signing. SW soc. Sci. Quart., 1956, 36, 385-390.

Bloch, V. L'etude objective du comportement des spectateurs. Rev. Int. de Filmol., 1952, 3, 221-222.

Boring, E. G. The role of theory in experimental psychology. Amer. J. Psychol., 1953, 66, 169-184. (Reprinted in E. G. Boring, History, psychology, and science. Watson, R. I., & Campbell, D. T. [Eds.], New York: Wiley, 1963, 210-225.)

Boring, E. G. The beginning and growth of measurement in psychology. Isis, 1961, 52, 238-257. (Reprinted in E. G. Boring, History, psychology, and science. Watson, R. I., & Campbell, D. T. [Eds.], New York: Wiley, 1963, 140-158.)

Boring, E. G. History, psychology and science. R. I. Watson, & D. T. Campbell (Eds.). New York: Wiley, 1963.

Boring, E.G., & Boring, M.D. Masters and pupils among the American psychologists. Amer. J. Psychol., 1948, 61, 527-534. (Reprinted in Boring, History, psychology, and science. Watson, R. I., & Campbell, D. T. [Eds.], New York: Wiley, 1963, 132-139.)

Brayfield, A. H., & Crockett, W. H. Employee attitudes and employee performance. Psychol. Bull., 1955, 52, 396-424.

Bridgman, P. W. The logic of modern physics. New York: Macmillan, 1927.

Brock, T. C., & Guidice, C. D. Stealing and temporal orientation. J. abnorm. soc. Psychol., 1963, 66, 91-94.

Brogden, H., & Taylor, E. The dollar criterion--applying the cost accounting concept to criterion construction. Personnel Psychol., 1950, 3, 133-154.

Brown, J. W. A new approach to the assessment of psychiatric therapies. Unpublished paper, 1960.

Brown, J. W. The use of the single case study with actuarial and indirect indices in psychiatric research. Unpublished paper, 1961.

Brozek, J. Recent developments in Soviet psychology. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.) Annu. Rev. Psychol., 1964, 15, 493-594.

Bugental, J. F. T. An investigation of the relationship of the conceptual matrix to the self-concept. Unpublished doctoral dissertation. Ohio State Univer., 1948.

Burchard, W. W. A study of attitudes towards the use of concealed devices in social science research. Soc. Forces, 1957, 36, 111.

Burchinal, L. G., & Chancellor, L. E. Ages at marriage, occupations of grooms and interreligious marriage rates. Soc. Forces, 1962, 40, 348-354.

Burchinal, L. G., & Chancellor, L. E. Survival rates among religiously homogamous and interreligious marriages. Soc. Forces, 1963, 41, 353-362.

Burchinal, L. G., & Kenkel, W.F. Religious identification and occupational status of Iowa grooms. *Amer. sociol. Rev.*, 1962, 27, 526-532.

Bureau of Applied Social Research. *A controlled study of the impact of anti-discrimination car cards.* New York: Bureau of Applied Social Research, Columbia Univer. and Dept. of Research and Evaluation. Anti-Defamation League of B'nai B'rith, 1953.

Burma, J. H. Self-tatooing among delinquents: a research note. *Sociol. soc. Res.*, 1959, 43, 341-345.

Burwen, R., & Campbell, D. T. The generality of attitudes toward authority and nonauthority figures. *J. abnorm. soc. Psychol.*, 1957, 54, 24-31.

Bush, G., & Londen, P. On the disappearance of knickers: hypothesis for the functional analysis of the psychology of clothing. *J. soc. Psychol.*, 1960, 51, 359-366.

Callahan, J. D., Morris, J.C., Seifried, S., Ulett, G. A., & Heusler, A. F. Objective measures in psychopharmacology: baseline observations. *Missouri Medicine*, 1960, 714-718.

Campbell, D. T. The informant in quantitative research. *Amer. J. Sociol.*, 1955, 60, 339-342.

Campbell, D. T. *Leadership and its effects upon the group.* Ohio Studies in Personnel, Research Monograph #83 Columbus: Bureau of Business Research, Ohio State Univer., 1956.

Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.*, 1957, 54, 297-312.

Campbell, D. T. Recommendations for APA test standards regarding construct trait or discriminant validity. *Amer. Psychologist*, 1960, 15, 546-553. (Reprinted in J. R. Braun (Ed.), *Clinical psychology in transition.* Cleveland: Howard Allen, 1961.

Campbell, D. T. The mutual methodological relevance of anthropology and psychology. In F. L. K. Hsu (Ed.), *Psychological anthropology approaches to culture and personality*. Homewood, Ill.: Dorsey Press, 1961. Pp. 333-352.

Campbell, D. T. From description to experimentation: interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in measuring change*. Madison, Wis.: Univer. Wisconsin Press, 1963, Pp. 212-242.

Campbell, D. T. Administrative experimentation, institutional records and nonreactive measures. In B. J. Chandler, E. F. Carlson, F. Bertolaet, C. Byerly, J. Lee, R. Sperber (Eds.), *Research Seminar on Teacher Education*, Report on Cooperative Research Project No. G-011 supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education, and Welfare. Duplicated, Northwestern University, August, 1963, pp. 75-120.

Campbell, D. T. Distinguishing differences of perception from failures of communication in cross-cultural studies. In F. S. C. Northrop, & H. H. Livingston (Eds.), *Cross-cultural understanding: epistemology in anthropology*. New York: Harper & Row, 1964, 308-336.

Campbell, D. T. Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *Probabilistic functionalism: the psychology of Egon Brunswik*. New York: Holt, Rinehart & Winston, 1965, in press.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 56, 81-105.

Campbell, D. T., & McCormack, T. H. Military experience and attitudes toward authority. Amer. J. Sociol., 1957, 62, 482-490.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching, Chicago: Rand McNally, 1963.

Campbell, N. R. An account of the principles of measurement and calculation. New York: Longmans, Green, 1928.

Campbell, N. R. Symposium: measurement and its importance for philosophy. Aristotelian Soc. Suppl., 1938, 17.

Cane, V. R., & Heim, A. W. The effects of repeated testing: III. Further experiments and general conclusions. Quart. J. exp. Psychol., 1950, 2, 182-195.

Cantril, H. Gauging public opinion. Princeton, N.J.: Princeton Univer. Press, 1944.

Caplow, T., & McGee, R. The academic marketplace. New York: Basic Books, 1958.

Capra, P. C., & Dittes, J. E. Birth order as a selective factor among volunteer subjects. J. abnorm. soc. Psychol., 1962, 64, 302.

Carlson, J., Cook, S. W., & Stromberg, E. L. Sex differences in conversation. J. appl. Psychol., 1936, 20, 727-735.

Carroll, P. F. Personal communication, 1962.

Chapman, L. J., & Bock, R. D. Components of variance due to acquiescence and content in the F-scale measure of authoritarianism. Psychol. Bull., 1958, 55, 328-333.

Chapple, E. D. Quantitative analysis of complex organizational systems. *Hum. Organization*, 1962, 21, 67-80.

Christensen, H. T. Cultural relativism and premarital sex norms. *Amer. sociol. Rev.*, 1960, 25, 31-39.

Clark, K. *America's psychologists*. Washington, D. C.: American Psychological Association, 1957.

Clark, W. H. A study of some of the factors leading to achievement and creativity with special reference to religious skepticism and belief. *J. soc. Psychol.*, 1955, 41, 57-69. (Abstracted in M. I. Stein, & S. J. Heinze [Eds.], *Creativity and the individual*. Glencoe, Ill.: Free Press, 1960.)

Cogley, J. Cited in McCahill, M. Parleys to evaluate Catholic status. *Chicago Sun-Times*, June 8, 1963.

Coleman, J., Katz, E., & Menzel, H. The diffusion of an innovation among physicians. *Sociometry*, 1957, 20, 253-270.

Coleman, R. P., & Neugarten, B. Social class in the city. In preparation.

Conrad, B. *The death of Manolete*. Cambridge: Houghton Mifflin, 1958.

Couch, A., & Keniston, K. Yeasayers and naysayers: agreeing response set as a personality variable. *J. abnorm. soc. Psychol.*, 1960, 60, 151-174.

Couch, A., & Keniston, K. Agreeing response set and social desirability. *J. abnorm. soc. Psychol.*, 1961, 62, 175-179.

Cox, C. M. The early mental traits of three hundred geniuses. In L. M. Terman (Ed.), *Genetic studies of genius*, Vol. II, Stanford Calif.: Stanford Univer. Press, 1926. (Abstracted in M. I. Stein, & S. J. Heinze, *Creativity and the individual*. Glencoe, Ill.: Free Press, 1960.)

Cox, G. H., & Marley, E. The estimation of motility during rest or sleep. _J. Neurol. Neurosurg. Psychiat._, 1959, 22, 57-60.

Craddick, R. A. Size of witch drawings as a function of time before, on, and after Halloween. _Amer. Psychologist_, 1962, 17, 307.

Cratty, J. Conformity behavior as a function of dress and race. Unpublished paper. Northwestern Univer., 1962.

Crespi, L. P. The interview effect in polling. _Publ. Opin. Quart._, 1948, 12, 99-111.

Cronbach, L. J. Response sets and test validity. _Educ. psychol. Measmt_, 1946, 6, 475-494.

Cronbach, L. J. Proposals leading to analytic treatment of social perception scores. In R. Tagiuri, & L. Petrullo (Eds.), _Person perception and interpersonal behavior_. Stanford, Calif.: Stanford Univer. Press, 1958. Pp. 353-379.

Crowald, R. H. Soviet grave markers indicate how buried rated with regime. _El Universal_ (Mexico City), August 14, 1964, 196, 12.

Davis, R. C. Physiological responses as a means of evaluating information. In A. D. Biderman, & H. Zimmer (Eds.), _The manipulation of human behavior_. New York: Wiley, 1961.

Dearborn, D. C., & Simon, H. A. Selective perception: a note on the departmental identification of executives. _Sociometry_, 1958, 21, 140-144.

DeCharms, R., & Moeller, G. Values expressed in American children's readers: 1800-1950. _J. abnorm. soc. Psychol._, 1962, 64, 136-142.

DeFleur, M. L., & Petranoff, R. M. A televised test of subliminal persuasion. _Publ. Opin. Quart._, 1959, 23, 168-180.

Dempsey, P. Liberalism-conservatism and party loyalty in the U. S. Senate. _J. soc. Psychol._, 1962, 56, 159-170.

Deutsch, M. An experimental study of the effects of cooperation and competition upon group process. Hum. Relat., 1949, 2, 199-231.

Dexter, L. A. What do congressmen hear? In N. Polsby, R. Dentler, & P. Smith (Eds.), Politics and social life. Boston: Houghton, 1963. Pp. 485-495.

Dexter, L. A. Communications--pressure, influence or education? In R. A. Bauer, I. de Sola Pool, & L. A. Dexter (Eds.), American business and public policy. New York: Atherton, 1963. (Reprinted in L. A. Dexter, & D.M. White [Eds.], People, society and mass communications. New York: Free Press, 1964. Pp. 394-409.)

Digman, J., & Tuttle, D. An interpretation of an election by means of obverse factor analysis. J. soc. Psychol., 1961, 53, 183-194.

Dittman, A. T., & Wynne, L. C. Linguistic techniques and the analysis of emotionality in interviews. J. abnorm. soc. Psychol., 1961, 63, 201-204.

Dollard, J., & Mowrer, O. H. A method for measuring tension in written documents. J. abnorm. soc. Psychol., 1947, 42, 3-32.

Donnelly, R. C., Goldstein, J., & Schwartz, R. D. Criminal law. New York: Free Press, 1962.

Doob, L. W. Communication in Africa. New Haven: Yale Univer. Press, 1961.

Dornbusch, S., & Hickman, L. Other-directedness in consumer goods advertising: a test of Riesman's historical theory. Soc. Forces, 1959, 38, 99-102.

Duncan, C. P. Personal communication, 1963.

Durand, J. Mortality estimates from Roman tombstone inscriptions. *Amer. J. Sociol.*, 1960, 65, 365-373.

Durkheim, E. *Suicide.* Trans. by J. A. Spaulding and G. Simpson, Glencoe, Ill.: Free Press, 1951.

Edwards, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden, 1957.

Edwards, A. L., & Walker, J. M. A note on the Couch and Keniston measure of agreement response set. *J. abnorm. soc. Psychol.*, 1961, 62, 173-174.

Ehrle, R.A., & Johnson, B. G. Psychologists and cartoonists. *Amer. Psychologist*, 1961, 16, 693-695.

Ekelblad, F. A. *The statistical method in business.* New York: Wiley, 1962.

Ellis, N. R., & Pryer, R. S. Quantification of gross bodily activity in children with severe neuropathology. *Amer. J. ment. Defic.*, 1959, 63, 1034-1037.

Enciso, J. *Design motifs of ancient Mexico.* New York: Dover, 1953.

Evan, W. M. Peer-group interaction and organizational socialization: a study of employee turnover. *Amer. sociol. Rev.*, 1963, 28, 429-435.

Exline, R. V. Explorations in the process of person perception: visual interaction in relation to competition, sex and affiliation. Technical Report #8 to ONR, Washington, D. C., Contract No. Nenr-2285 (02), May 15, 1962.

Fairbanks, H. The quantitative differentiation of samples of spoken language. *Psychol. Monogr.*, 1944, 57, 19-38.

Fantz, R. L. A method for studying depth perception in infants under six months of age. *Psychol. Rec.*, 1961, 11, 27-32. (a)

Fantz, R. L. Origin of form perception. Scient. American, May, 1961. (b)

Fantz, R. L., Ordy, J. M., & Udelf, M.S. Maturation of pattern vision in infants during the first six months. J. comp. physiol. Psychol., 1962, 55, 907-917.

Fantz, R. L. Pattern vision in newborn infants. Science, 1963, 140, 296-297.

Farris, C. D. A method of determining ideological groupings in the Congress. J. Politics, 1958, 20, 308-338.

Feshbach, S., & Feshbach, N. Influence of the stimulus object upon the complementary and supplementary projection of fear. J. abnorm. soc. Psychol., 1963, 66, 498-502.

Fiedler, F.E.,The nature of teamwork. Discovery, February, 1962.

Fiedler, F.E., Dodge, J. S., Jones, R. E., & Hutchins, E. B. Interrelations among measures of personality adjustment in nonclinical population. J. abnorm. soc. Psychol., 1958, 56, 345-351.

Fisher, I. The making of index numbers. New York: Houghton-Mifflin, 1923.

Fiske, D. W. Values, theory and the criterion problem. Personnel Psychol., 1951, 4, 93-98.

Flagler, J. M. Profiles: student of the spontaneous. New Yorker, 1960, 36, 59-92.

Flugel, J. C. Psychology of clothes. London: Hogarth, 1930.

Foote, E. Pupil dilation--new measurement of ad's effectiveness. Advertising Age, 1962.

Foshee, J. G. Studies in activity level: I. Simple and complex task performances in defectives. Amer. J. ment. Defic., 1958, 62, 882-886.

Fowler, E. M. Help-wanted ads show sharp rise. *New York Times*, May 13, 1962, p. 1.

Franzen, R. Scaling responses to graded opportunities. *Publ. Opin. Quart.*, 1950, 14, 484-490.

Freed, A., Chandler, P. J., Mouton, J. S., & Blake, R. R. Stimulus and background factors in sign violation. *J. Pers.*, 1955, 23, 499.

Freeman, L. C., & Ataov, T. Invalidity of indirect and direct measures of attitude toward cheating. *J. Pers.*, 1960, 28, 443-447.

French, E. G. Some characteristics of achievement motivation. *J. exp. Psychol.*, 1955, 50, 232-236.

French, N. R., Carter, C. W., & Koenig, W. The words and sounds of telephone conversations. *Bell Sys. tech. J.*, 1930, 9, 290-324.

Freud, S. *Psychopathology of everyday life*. London: Unwin, 1920.

Fry, C. L. The religious affiliations of American leaders. *Scient. Monthly*, 1933, 36, 241-249. (Abstracted in M. I. Stein & S. J. Heinze, *Creativity and the individual*. Glencoe, Ill.: Free Press, 1960.)

Gabriele, C. T. The recording of audience reactions by infrared photography, Technical Report: NAVTRADEVCEN 269-7-56, 1956.

Gage, N. L., & Shimberg, B. Measuring senatorial progressivism. *J. abnorm. soc. Psychol.*, 1949, 44, 112-117.

Galton, F. Hereditary genius, New York: D. Appleton, 1870. (Abstracted in M. I. Stein & S. J. Heinze, *Creativity and the individual*. Glencoe, Ill.: Free Press, 1960.)

Galton, F. Measurement of character. *Fortnightly Rev.*, 1884, 36, 179-185.

Garner, W. R. Context effects and the validity of loudness scales. J. exp. Psychol., 1954, 48, 218-224.

Garner, W. R. Hake, H. W., & Eriksen, C. W. Operationism and the concept of perception. Psychol. Rev., 1956, 63, 149-159.

Gearing, F. The response to a cultural precept among migrants from Bronzeville to Hyde Park. Unpublished master's thesis, Department of Anthropology, Univer. Chicago, June, 1952.

Ghiselli, E. E., & Brown, C. W. Personnel and industrial psychology. New York: McGraw-Hill, 1955.

Goncourt, Edmond de. The Goncourt Journals 1851-1870.(Edited and translated from the Journal of Edmond & Jules de Goncourt. New York: Doubleday, Doran, 1937.)

Good, C. V., & Scates, D. E. Methods of research, educational, psychological, sociological. New York: Appleton-Century-Crofts, 1954.

Gordon, T. The development of a method of evaluating flying skill. Personnel Psychol., 1950, 3, 71-84.

Gore, P. M., & Rotter, J. B. A personality correlate of social action. J. Pers., 1963, 31, 58-64.

Gosnell, H. F. Getting out the vote: an experiment in the stimulation of voting. Chicago: Univer. Chicago, 1927.

Gottschalk, L. A., & Gleser, G. C. An analysis of the verbal content of suicide notes. Brit. J. med. Psychol., 1960, 33, 195-204.

Gould, J. New York Times, 1951. Cited in I. Doig, Unpublished master's thesis, Northwestern Univer., 1962.

Grace, H., & Tandy, M. Delegate communication as an index of group tension. J. soc. Psychol., 1957, 45, 93-97.

Gratiot-Alphandery, H. L'enfant et le film. Rev. Int. de Filmol., 1951, 2, 171-172. (a)

Gratiot-Alphandery, H. Jeunes spectateurs. Rev. Int. de Filmol., 1951, 2, 257-263.

Green, E. Judicial attitudes in sentencing. New York: St. Martin's Press, 1961.

Green, H. B., & Knapp, R. H. Time judgment, aesthetic preference, and need for achievement. J. abnorm. soc. Psychol., 1959, 58, 141-142.

Griffith, R. M. Odds adjustments by American horse race bettors. Amer. J. Psychol., 1949, 62, 290-294.

Grinder, R. E. New techniques for research in children's temptation behavior. Child Develpm., 1961, 32, 679-688.

Grinder, R. E. Parental child learning practices, conscience, and resistance to temptation of sixth grade children. Child Develpm., 1962, 33, 803-820.

Grinder, R. E., & McMichael, R. E. Cultural influence on conscience development: resistance to temptation and guilt among Samoans and American caucasians. J. abnorm. soc. Psychol., 1963, 66, 503-507.

Grusky, O. Organizational goals and the behavior of informal leaders. Amer. J. Sociol., 1959, 65, 59-67.

Grusky, O. Managerial succession and organizational effectiveness. Amer. J. Sociol., 1963, 69, 21-31.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954.

Guilford, J. P. The relation of intellectual factors to creative thinking in science. In C. Taylor (Ed.), The 1955 University of Utah research conference on the identification of creative scientific talent. Salt Lake City: Univer. Utah Press, 1956. Pp. 69-95.

Gujon, R. M. Criterion measurement and personnel judgments. Personnel Psychol., 1961, 14, 141-149.

Gullahorn, J., & Strauss, G. The field worker in union research. In R. N. Adams, & J. J. Preiss (Eds.), Human organization research. Homewood, Ill.: Dorsey Press, 1960. Pp. 153-165.

Gump, R. Jade: stone of heaven. New York: Doubleday, 1962.

Gusfield, J. R. Field work reciprocities in studying a social movement. In R. N. Adams, & J. J. Preiss (Eds.), Human organization research. Homewood, Ill.: Dorsey Press, 1960.

Haggard, E. A., Brekstad, A., & Skard, A. G. On the reliability of the anamnestic interview. J. abnorm. soc. Psychol., 1960, 61, 311-318.

Halbwachs, M. Les Causes de Suicide. Paris: Felix Alcan, 1930.

Hall, G. S. Chilrens' lies. Amer. J. Psychol., 1891, 3, 59-70.

Hamburger, P. Peeping Funt. New Yorker, 1950, 25, 72-73.

Hamilton, T. Social optimism in American protestanism. Publ. Opin. Quart., 1942, 6, 280-283.

Hansen, A. H. Cycles of prosperity and depression in The United States. Univer. Wisconsin studies in social sciences and history. Madison, Wis., 1921.

Hanson, N. R. Patterns of discovery. Cambridge, England: Cambridge Univer. Press, 1958.

Hartshorne, H., & May, M. A. *Studies in the nature of character* - I. New York: Macmillan, 1928.

Hartshorne, H., & May, M. A. *Studies in service and self-control* - II. New York: Macmillan, 1929.

Harvey, J. The content of best-selling novels. *Publ. Opin. Quart.*, 1953, 17, 91-114.

Helson, H., Blake, R. R., & Mouton, J. S. Petition-signing as adjustment to situational and personal factors. *J. soc. Psychol.*, 1958, 48, 3-10.

Henle, M., & Hubble, M. B. Egocentricity in adult conversation. *J. soc. Psychol.*, 1938, 9, 227-234.

Henry, H. *Motivation research: its practice and uses for advertising, marketing, and other business purposes.* London: Crosby Lockwood, 1958.

Herbiniere-Lebert, S. Pourquoi et comment nous avons fait "Mains Blanches": premieres experiences avec un film educatif realise specialement pour les moins de sept ans. *Rev. Int. de Filmol.*, 1951, 2, 247-255.

Hess, E. H., & Polt, J. M. Pupil size as related to interest value of visual stimuli. *Science*, 1960, 132, 349-350.

Heusler, A., Ulett, G., & Blasques, J. Noise-level index: an objective measurement of the effect of drugs on the psychomotor activity of patients. *J. Neuropsychiat.*, 1959, Vol. 1, No. 1.

Heusler, A. F., Ulett, G. A., & Callahan, J. D. Comparative EEG studies of tranquilizing drugs. Research Laboratories of the St. Louis State Hospital, St. Louis, Mo. Presented at Pan-American Medical Congress. Mexico City, May 3, 1960.

Heyns, R., & Lippitt, R. Systematic observational techniques. In G. Lindzey (Ed.), Handbook of social psychology. Vol. I. Cambridge: Addison-Wesley, 1954. Pp. 370-404.

Hildum, D. C., & Brown, R. W. Verbal reinforcement and interviewer bias. J. abnorm. soc. Psychol., 1956, 53, 108-111.

Hillebrandt, R. H. Panel design and time-series analysis. Unpublished master's thesis, Northwestern Univer., 1962.

Holmes, L. D. Ta'u: Stability and change in a Samoan village. Wellington, New Zealand: Polynesian Society Reprint No. 7, 1958.

Houseman, E.E., & Lipstein, B. Observation and audit techniques for measuring retail sales. Agricultural Economics, 1960, 12, 61-72.

Hovland, C. I. Effects of the mass media of communication. In G. Lindzey (Ed.), Handbook of social psychology. Vol. II. Cambridge, Mass.: Addison Wesley, 1954.

Hovland, C. I., Lumsdaine, A. A., & Sheffield, F. D. Experiments on mass communication. Princeton, N.J.: Princeton Univer. Press, 1949.

Howells, L. T., & Becker, S. W. Seating arrangement and leadership emergence. J. abnorm. soc. Psychol., 1962, 64, 148-150.

Humphreys, L. G. Note on the multitrait-multimethod matrix. Psychol. Bull., 1960, 57, 86-88.

Hyman, H. H., Cobb, W. J., Feldman, J. J., Hart, C. W., & Stember, C. H. Interviewing in social research. Chicago: Univer. Chicago Press, 1954.

Ianni, F. A. Residential and occupational mobility as indices of the acculturation of an ethnic group. Soc. Forces, 1957-58, 36, 65-72.

Imanishi, K. Social organization of subhuman primates in their natural habitat. *Curr. Anthropol.*, 1960, 1, 393-407.

Jackson, D. N., & Messick, S. J. A note on "ethnocentrism" and acquiescent response sets. *J. abnorm. soc. Psychol.*, 1957, 54, 132-134.

Jacques, E. *Measurement of responsibility*. Cambridge: Harvard Univer. Press, 1956. (Book Review in *Publ. Opin. Quart.*, 1957-58, 21, 329.)

James, J. A preliminary study of the size determinant in small group interaction. *Amer. Sociol. Rev.*, 1951, 16, 474-477.

James, R. W. A technique for describing community structure through newspaper analysis. *Soc. Forces*, 1958, 37, 102-109.

Janis, I. L., & Hovland, C. I. An overview of persuasability research. In I. L. Janis, & C. I. Hovland (Eds.), *Personality and persuasability*. New Haven: Yale Univer. Press, 1959. Pp. 1-26.

Janowitz, M. Inferences about propaganda impact from textual and documentary analysis. In W. E. Daugherty,& M. Janowitz (Eds.), *A psychological warfare casebook*. Baltimore: Johns Hopkins Press, 1958, 732-735.

Jay, R., & Copes, J. Seniority and criterion measures of job proficiency. *J. appl. Psychol.*, 1957, 41, 58-60.

Jones, R. W. Progressivism in Illinois communities as measured by library services. *Transactions of the Illinois State Academy of Science*, 1960, 53, 166-172.

Jung, A. F. Price variations among automobile dealers in Chicago, Illinois. *J. Bus.*, 1959, 32, 315-326.

Jung, A. F. Prices of Falcon and Corvair cars in Chicago and selected cities. *J. Bus.*, 1960, 33, 121-126.

Jung, A. F. Impact of the compact cars on new-car prices. *J. Bus.*, 1961, 34, 167-182.

Jung, A. F. Impact of the compact cars on new-car prices: a reappraisal. *J. Bus.*, 1962, 35, 70-76.

Jung, A. F. Dealer pricing practices and finance charges for new mobile homes. *J. Bus.*, 1963, 36, 430-439.

Kahn, R. L., & Cannell, C. F. *The dynamics of interviewing: theory, technique and cases.* New York: Wiley, 1957.

Kaminski, G., & Osterkamp, U. Untersuchungen über die Topologie sozialer Handlungsfelder. *Z. exp. angew. Psychol.*, 1962, 9, 417-451.

Kane, F. Clothing worn by out-patients to interviews. *Psychiat. Communications*, 1958, 1, #2.

Kane, F. Clothing worn by an out-patient: a case study. *Psychiat. Communications*, 1959, 2, #2.

Kane, F. The meaning of the form of clothing. *Psychiat. Communi*cations, 1962, 5, #1.

Kanfer, F. H. Verbal rate, eyeblink and content in structured psychiatric interviews. *J. abnorm. soc. Psychol.*, 1960, 61, 341-347.

Kappel, J. W. Book clubs and the evaluation of books. *Publ. Opin. Quart.*, 1948, 12, 243-252.

Katz, D. Do interviewers bias poll results? *Publ. Opin. Quart.*, 1942, 2, 248-268.

Kavanau, J. L. Behavior: confinement, adaptation, and compulsory regimes in laboratory studies. *Science*, 1964, 143, 490.

Kendall, L. M. Paper presented at the symposium: "The Hidden Variance: What does it Measure?" American Psychological Association annual convention, Philadelphia, Sept. 3, 1963. (Amer. Psychologist, 1963, 18, 452)

Kimbrell, D. L., & Blake, R. R. Motivational factors in the violation of a prohibition. J. abnorm. soc. Psychol., 1958, 56, 132-133.

Kinsey, A. C., Pomeroy, W. B., Martin, C. E., & Gebhard, P. H. Sexual behavior in the human female. Philadelphia: W. B. Saunders, 1953.

Kirk, P. L. Criminalistics. Science, 1963, 140, 367-370.

Kitsuse, J. I., & Cicourel, A. V. A note on the uses of official statistics. Soc. Probl., 1963, 11, 131-139.

Klapper, J. What we know about the effect of mass communication: the brink of hope. Publ. Opin. Quart., 1957-58, 21, 453-474.

Klapper, J. The effects of the mass media. Glencoe, Ill.: Free Press, 1960.

Knox, J. B. Absenteeism and turnover in an Argentine factory. Amer. sociol. Rev., 1961, 26, 424-428.

Kort, F. Predicting Supreme Court decisions mathematically: a quantitative analysis of "right to counsel" cases. Amer. Pol.Sci. Rev., 1957, 51, 1-12.

Kort, F. Reply to Fisher's "Mathematical analysis of Supreme Court decisions." Amer. Pol. Sci. Rev., 1958, 52, 339-348.

Kramer, E. Judgment of personal characteristics and emotions from nonverbal properties of speech. Psychol. Bull., 1963, 60, 408-420.

Kramer, E. Elimination of verbal cues in judgments of emotion from voice. J. abnorm. soc. Psychol., 1964, 68, 390-396.

Krasner, L. Studies of the conditioning of verbal behavior. Psychol. Bull., 1958, 55, 148-170.

Kretsinger, E. A. An experimental study of gross bodily movement as an index to audience interest. Speech Monogr., 1952, 19, 244-248.

Kretsinger, E. A. An experimental study of restiveness in preschool educational television audiences. Speech Monogr., 1959, 26, 72-77.

Krislov, S. Yale Law Journal, 1963, 72, 694.

Krout, M. H. Major aspects of personality. Chicago: College Press, 1933.

Krout, M. H. Autistic gestures: an experimental study in symbolic movement. Psychol. Monogr., 1935, 46, No. 4 (Whole No. 208).

Krout, M. H. Further studies on the relation of personality and gesture: a nosological analysis of autistic gestures. J. exp. Psychol., 1937, 20, 279-287.

Krout, M. H. Gestures and attitudes: an experimental study of the verbal equivalents and other characteristics of a selected group of manual autistic gestures. Unpublished doctoral dissertation, Univer. of Chicago, 1951.

Krout, M. H. An experimental attempt to determine the significance of unconscious manual symbolic movements. J. gen. Psychol., 1954, 51, 121-152 (a)

Krout, M. H. An experimental attempt to produce unconscious manual symbolic movements. J. gen. Psychol., 1954, 51, 93-120.

Kuhn, T. The structure of scientific revolutions. Chicago: Univer. of Chicago Press, 1962.

Lander, B. Towards an understanding of juvenile delinquency. New York: Columbia Univer. Press, 1954.

Landis, C. National differences in conversation. J. abnorm. soc. Psychol., 1927, 21, 354-357.

Landis, C., & Hunt, W.A. The startle pattern. New York: Farrar & Rinehart, 1939.

Landis, M. H., & Burtt, H. E. A study of conversations. J. comp. Psychol., 1924, 4, 81-89.

Lang, K., & Lang, G. E. The mass media and voting. In E. Burdick, & A. J. Brodbeck (Eds.), American voting behavior. Glencoe, Ill.: Free Press, 1959.

Lang, K., & Lang, G.E. Decisions for Christ: Billy Graham in New York City. In Villich et al. (Eds.), Identity and anxiety. Glencoe, Ill.: Free Press, 1960.

LaPiere, R. T. Attitudes vs. actions. Soc. Forces, 1934, 13, 230-237.

Lea, T. The brave bulls. Boston: Little, Brown, 1949.

Lefkowitz, M., Blake, R. R., & Mouton, J. S. Status factors in pedestrian violation of traffic signals. J. abnorm. soc. Psychol., 1955, 51, 704-706.

Lehman, H. C., & Witty, P. A. Scientific eminence and church membership. Scient. Monthly, 1931, 36, 544-549.
(Abstracted in M. I., Stein, & S. J. Heinze, Creativity and the individual. Glencoe, Ill.: Free Press, 1960.)

Leipold, W. D. Psychological distance in a dyadic interview as a function of introversion-extraversion, anxiety, social desirability and stress. Unpublished doctoral dissertation, Univer. North Dakota, 1963.

Lenski, G. E., & Leggett, J. C. Caste, class, and deference in the research interview. Amer. J. Soc., 1960, 65, 463-467.

Leroy-Boussion, A. Etude du comportement emotional enfantin au cours de la projection d'un film comique. Rev. Int. de Filmol., 1954, 5, 105-123.

Lewis, O. The children of Sanchez. New York: Random House, 1961.

Libby, W. I. Accuracy of radio-carbon dates. Science, 140, 1963, 278-280.

Lippmann, W. The public philosophy. New York: New American Library, 1955.

Lodge, G.T. Pilot stature in relation to cockpit size: a hidden factor in Navy jet aircraft accidents. Amer. Psychologist, 1963, 17, 468.

Lombroso, C. The man of genius. London: Walter Scott, 1891. (Abstracted in M. I. Stein, & S. J. Heinze, Creativity and the individual. Glencoe, Ill.: Free Press, 1960.

Loomis, C. P. Political and occupational changes in a Hanoverian village, Germany. Sociometry, 1946, 9, 316-333.

Lucas, D. B., & Britt, S. H. Advertising psychology and research. New York: McGraw-Hill, 1950.

Lucas, D. B., & Britt, S. H. Measuring advertising effectiveness. New York: McGraw-Hill, 1963.

Lustig, N. I. The relationships between demographic characteristics and pro-integration vote of white precincts in a metropolitan southern community. Soc. Forces, 1962, 40, 205-208.

Lyle, H. M. An experimental study of certain aspects of the electromagnetic movement meter as a criterion to audience attention. Speech Monogr., 1953, 20, 126.

Mabie, E. A study of the conversation of first-grade pupils during free play periods. *J. educ. Res.*, 1931, 24, 135-138.

Mabley, J. Mabley's report. *Chicago American*, January 22, 1963.

McCarthy, D. A comparison of children's language in different situations and its relation to personality traits. *J. genet. Psychol.*, 1929, 36, 583-591.

Maccoby, E. E. Developmental psychology. In P. R. Farnsworth, O. McNemar, & Q. McNemar. *Annu. Rev.*, 1964, 15, 203-250.

Maccoby, N. Experiments in nonverbal cognitive feedback to teachers. Research proposal submitted to the Department of Health, Education, and Welfare, 1963.

McGrath, J. E. The influence of positive interpersonal relations on adjustment and effectiveness in rifle teams. *J. abnorm. soc. Psychol.*, 1962, 65, 365-375.

McGraw, M., & Molloy, L. B. *The pediatric anamnesis: inaccuracies in* eliciting developmental data. *Child Develpm.*, 1941, 12, 255-265.

MacKinney, A. C. What should ratings rate? *Personnel*, 1960, 37, 75-78.

MacRae, D. The role of the state legislator in Massachusetts. *Amer. sociol. Rev.*, 1954, 19, 185-194. (a)

MacRae, D. Some underlying variables in legislative roll call votes. *Publ. Opin. Quart.*, 1954, 18, 191-196. (b)

MacRae, D., & MacRae, E. Legislators' social status and their votes. *Amer. J. Soc.*, 1961, 66, 599-603.

Mahl, G. Disturbances and silences in the patient's speech in psychotherapy. *J. abnorm. soc. Psychol.*, 1956, 53, 1-15.

Manago, B. R. Mad: out of the comics rack and into satire. *Add One*, 1962, 1 (2), 41-46.

Marsh, R. M. Formal organization and promotion in a pre-industrial society. *Amer. sociol. Rev.*, 1961, 26, 547-556.

Martin, P. I call on the candid camera man. Sat. Eve. Post, 1961, 234, 26-27.

Matarazzo, J. D. Control of interview behavior. Paper presented at annual meeting of the American Psychological Association, 1962. (a)

Matarazzo, J. D. Prescribed behavior therapy: suggestions from noncontent interview research. In A. J. Bachrach (Ed.), Experimental foundations of clinical psychology. New York: Basic Books, 1962, 471-509. (b)

Matarazzo, J. D., Weitman, M., Saslow, G., & Wiens, A. N. Interviewer influence on duration of interviewee speech. J. Verbal Learn. Verbal Behav., 1963, 1, 451-458

Matarazzo, J. D., Wiens, A. N., Saslow, G., Dunham, R.M., & Voas, R. B. Speech durations of astronaut and ground communicator. Science, 1964, 143, 148-150.

Matthews, T. S. What makes news. The Atlantic, December, 1957, 200, 80-83.

May, M. A., & Hartshorne, H. Objective methods of measuring character Pedagogical Seminary, 1925, 32, 45-67.

Meadow, A., Greenblatt, M., Levine, J., & Solomon, H. The discomfort-relief quotient as a measure of tension and adjustment. J. abnorm. soc. Psychol., 1952, 47, 658-661

Mechanic, D., & Volkart, E. H. Stress, illness behavior and the sick role. Amer. sociol. Rev., 1961, 26, 51-58.

Melbin, M. Organization practice and individual behavior: absenteeism among psychiatric aides. Amer. sociol. Rev., 1961, 26, 14-23.

Melton, A. W. Some behavior characteristics of museum visitors. Psychol. Bull., 1933, 30, 720-721. (a)

Melton, A. W. Studies of installation at the Pennsylvania Museum of Art. *Museum News*, 1933, 11, 508. (b)

Melton, A. W. Problems of installation in museums of art. *Studies in museum education*, The American Association of Museums, Washington, D. C., 1935.

Melton, A. W. Distribution of attention in galleries in a museum of science and industry. *Museum News*, 1936.

Melton, A. W., Feldman, N. G., & Mason, C. W. *Experimental studies of the education of children in a museum of science*. Publication of the American Association of Museums, New Series, No. 15, 1936.

Middleton, R. Fertility values in American magazine fiction: 1916-1956. *Publ. Opin. Quart.*, 1960, 24, 139-143.

Miller, G. A. Population, distance and the circulation of information. *Amer. J. Psychol.*, 1947, 60, 276-284.

Mills, F. C. *The behavior of prices.* New York: National Bureau of Economic Research, 1927.

Mindak, W. A., Neibergs, A., & Anderson, A. Economic effects of the Minneapolis newspaper strike. *Journalism Quart.*, 1963, 40, 213-218.

Mitchell, W. C. *The making and using of index numbers*. Part I of *Index Numbers of Wholesale Prices in the U. S. and Foreign Countries*. Bull. No. 284, Bureau of Labor Statistics, 1921.

Moore, H. T. Laboratory tests of anger, fear, and sex interest. *Amer. J. Psychol.*, 1917, 28, 390-395.

Moore, H. T. Further data concerning sex differences. *J. abnorm. soc. Psychol.*, 1922, 17, 210-214.

Moore, U., & Callahan, C. *Law and learning theory: a study in legal control*. New Haven: Yale Law Journal Co., 1943.

Morgan, E. M. *Basic problems of evidence*. New York: Joint committee on continuing legal education of the American Law Institute and the American Bar Association, 1963.

Mosteller, F., & Wallace, D. L. Inference in an authorship problem: a comparative study of discrimination methods applied to the authorship of *The Federalist Papers*. Paper read before Amer. Stat. Association, Biometric Society and Inst. Math. Stat., Minneapolis, 1962.

Mudgett, B. D. *Index numbers*. New York: Wiley, 1951.

Murphy, G., & Murphy, L. Soviet life and Soviet psychology. In R. A. Bauer (Ed.), *Some views on Soviet psychology*. Washington: American Psychological Association, 1962. Pp. 253-276.

Murray, E. J., & Cohen, M. Mental illness, milieu therapy, and social organization in ward groups. *J. abnorm. soc. Psychol.*, 1959, 58, 48-54.

Nagel, S. Ethnic affiliations and judicial propensities. *J. Politics*, 1962, 24, 92.

Naroll, R. The preliminary index of social development. *Amer. Anthropologist*, 1956, 58, 687-715.

Naroll, R. Controlling data quality. In/Series *Research in social psychology*. Symposia Studies Series, No. 4, September 1960, 7-12.

Naroll, R. Two solutions to Galton's problems. *Phil. Sci.*, 1961, 28, 15-39.

Naroll, R. *Data quality control.* Glencoe, Ill.: Free Press, 1962.

Naroll, R., & Naroll, F. On bias of exotic data. *Man,* 1963, 25, 24-26.

NASA Manned Spacecraft Center. *Results of the first United States manned orbital space flight,* February 20, 1962. Washington: Government Printing Office, 1962. (a)

NASA Manned Spacecraft Center. *Results of the second United States manned orbital space flight,* May 24, 1962. Government Printing Office, 1962. (b)

National Advertising Company. *Shopping center research study.* Bedford Park, Ill., 1963.

Nixon, H. K. Attention and interest in advertising. *Arch. Psychol.,* 1924, 11, 1-65.

O'Connor, N. (Ed.) *Recent Soviet psychology.* Kish, C., & Asher, H. (trans.). New York: Liveright, 1961.

Olson, W. C. *The measurement of nervous habits in children.* Minneapolis: Univer. of Minnesota, 1929.

Osgood, C. E. *Method and theory in experimental psychology.* New York: Oxford Univer. Press, 1953.

Osgood, C. E., & Walker, E. Motivation and language behavior: a content analysis of suicide notes. *J. abnorm. soc. Psychol.,* 1959, 59, 58-67.

OSS Assessment Staff, *Assessment of men.* New York: Rinehart, 1948.

Parker, E. B. The effects of television on public library circulation. Unpublished, read at the annual convention of the Pacific Chapter, American Association for Public Opinion Research, Pacific Grove, California, January, 1963.

Perrine, M., & Wessman, A. W. Disguised public opinion interviewing with small samples. Publ. Opin. Quart., 1954, 18, 92-96.

Phillips, R. H. Miami goes Latin under Cuban tide. New York Times, March 18, 1962.

Polansky, N., Freeman, W., Horowitz, M., Irwin, L. Paponia, N., Rapaport, D., & Whaley, F. Problems of interpersonal relations in research on groups. Hum. Relat., 1949, 2, 281-291.

Politz Media Studies. Ad Page Exposure in Four Magazines. New York: Alfred Politz, 1957.

Politz Media Studies. A study of outside transit poster exposure. New York: Alfred Politz, 1959.

Popper, K. Logic der Forschung. Wien, Springer, 1935.

Popper, K. . The logic of scientific discovery. New York: Basic Books, 1959.

Popper, K. . Conjectures & refutations. New York: Basic Books, 1962.

Pyles, M. K., Stolz, H. R., & Macfarlane, J. W. The accuracy of mothers' reports on birth and developmental data. Child Develpm., 1935, 6, 165-176.

Quine, W. V. From a logical point of view. Cambridge, Mass.: Harvard, 1953.

Rashkis, H., & Wallace, A. F. C. The reciprocal effect. AMA Arch gen. Psychiat., 1959, 1, 489-498.

Riesman, D. Orbits of tolerance, interviewers and elites. Publ. Opin. Quart., 1956, 20, 49-73.

Riesman, D. The state of communication research. Publ. Opin. Quart., 1959, 23, 10-13. (Part of an article by B. Berelson, pp. 1-15.)

Riesman, D., & Ehrlich, J. Age and authority in the interview. *Publ. Opin. Quart.*, 1961, 25, 39-56.

Riker, W., & Niemi, D. The stability of coalitions on roll calls in the house of representatives. *Amer. pol. sci. Rev.*, 1962, 56, 58-65.

Riley, M. W. *Sociological research.* New York: Harcourt, Brace & World, 1963.

Robbins, L. C. The accuracy of parental recall of aspects of child development and of child rearing practices. *J. abnorm. soc. Psychol.*, 1963, 66, 261.

Robins, L. N., Hyman, H., & O'Neal, P. The interaction of social class and deviant behavior. *Amer. sociol., Rev.*, 1962, 27, 480-492.

Robinson, D., & Rohde, S. Two experiments with an anti-semiticism poll. *J. abnorm. soc. Psychol.*, 1946, 41, 136-144.

Robinson, E. S. The behavior of the museum visitor. Publication of the American Association of Museums, New Series, No. 5, 1928.

Roens, B. B. New findings from Scott's special advertising research study. In *Proceedings: 7th Annual Conference, Advertising Research Foundation.* New York: Advertising Research Foundation, 1961.

Rosenbaum, M. E. The effect of stimulus and background factors on the volunteering response. *J. abnorm. soc. Psychol.*, 1956, 53, 118-121.

Rosenbaum, M. E., & Blake, R. R. Volunteering as a function of field structure. *J. abnorm. soc. Psychol.*, 1955, 50, 193-196.

Rosenthal, A. M. Japan, famous for politeness, has a less courteous side, too. *New York Times*, February 25, 1962.

Rosenthal, R. On the social psychology of the psychological experiment. *Amer. Sci.*, 1963, 51, 268-283.

Rosenthal, R. Experimenter outcome-orientation and the results of the psychological experiment. Psychol. Bull., 1964, 61, 405-412.

Rotter, J. B., Liverant, S., & Crowne, D. P. The growth and extinction of expectancies in chance, controlled and skilled tasks. J. Psychol., 1961, 52, 161-177.

Ruesch, J., & Kees, W. Nonverbal communication: notes on the visual perception of human relations. Berkeley: Univer. of Calif. Press, 1956.

Rush, C. A factorial study of sales criteria. Personnel Psychol., 1953, 6, 9-24.

Salzinger, K. A method of analysis of the process of verbal communication between a group of emotionally disturbed adolescents and their friends and relatives. J. soc. Psychol., 1958, 47, 39-53.

Sawyer, H. G. The meaning of numbers, 1961, speech before the American Association of Advertising Agencies.

Schacter, S., & Hall, R. Group derived restraints and audience persuasion. Hum. Relat., 1952, 5, 397-406.

Schanck, R. L., & Goodman, C. Reactions to propaganda on both sides of a controversial issue. Publ. Opin. Quart., 1939, 3, 107-112.

Shneidman, E. S., & Farberow, N. L. Some comparisons between genuine and simulated suicide notes in terms of Mowrer's concepts of discomfort and relief. J. gen. Psychol., 1957, 56, 251-256.

Schubert, G. Quantitative analysis of judicial behavior. Glencoe, Ill.: Free Press, 1959.

Schubert, G. Judicial decision-making. New York: Free Press of Glencoe, 1963.

Schulman, J. Personal communication, 1962.

Schwartz, M. S., & Stanton, A. H. A social psychological study of incontinence. *Psychiatry*, 1950, 13, 399-416.

Schwartz, R. D. Field experimentation in sociolegal research. *J. legal Education*, 1961, 13, 401-410.

Schwartz, R. D., & Skolnick, J. H. Two studies of legal stigma. *Social problems*, 1962, 10, 133-142.

Sebald, H. Studying national character through comparative content analysis. *Soc. Forces*, 1962, 40, 318-322.

Sechrest, L., & Wallace, J. Personal communication, 1964.

Selltiz, C., Jahoda, M., Deutsch, M., & Cook, S. W. *Research methods in social relations*. New York: Holt, Rinehart & Winston, 1959.

Severin, D. The predictability of various kinds of criteria. *Personnel Psychol.*, 1952, 5, 93-104.

Shadegg, S. C. *How to win an election*. New York: Toplinger, 1964.

Shepard, H. R., & Blake, R. R. Changing behavior through cognitive change. *Hum. Organization*, 1962, 21, 88-92.

Siersted, E., & Hansen, H. L. Reaction des petits enfants au cinema: resume d'une serie d'observations faites au Danemark. *Revue Int. de Filmol.*, 1951, 2 (7-8), 241-245.

Singh, P. N., & Huang, S. C. Some socio-cultural and psychological determinants of advertising in India: a comparative study. *J. soc. Psychol.*, 1962, 57, 113-121.

Sleeper, C. B. Samplings of leisure-time conversations to find sex differences in drives. Unpublished, cited in Murphy and Murphy. New York: Harper, 1931.

Sletto, R. F. *A construction of personality scales by the criterion of interval consistency*. Hanover, N. H.: Soc. Press, 1937.

Smedslund, J. Educational psychology. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.), *Annu. Rev. Psychol.*, 1964, 15, 251-276.

Smith, H. T. A comparison of interview and observation methods of mother behavior. *J. abnorm. soc. Psychol.*, 1958, 57, 278-282.

Snyder, E. C. Uncertainty and the Supreme Court's decisions. *Amer. J. Soc.*, 1959, 65, 241-245.

Solomon, R. L. An extension of control group design. *Psychol. Bull.*, 1949, 46, 137-150.

Sommer, R. Studies in personal space. *Sociometry*, 1959, 22, 247-260.

Sommer, R. Personal space. *Canad. Architect*, 1960. Pp. 76-80.

Sommer, R. Leadership and group geography. *Sociometry*, 1961, 24, 99-110.

Sommer, R. The distance for confortable conversations: further study. *Sociometry*, 1962, 25, 111-116.

Spiegel, D.E., & Neuringer, C. Role of dread in suicidal behavior. *J. abnorm. soc. Psychol.*, 1963, 66, 507-511.

Spohn, H. E., & Hunter, . Reported in J. W. Brown. The use of the single study with actuarial and indirect indices in psychiatric research, 1961.

Stechler, G. Newborn attention as affected by medication during labor. *Science*, 1964, 144, 315-317.

Stein, M. I., & Heinze, S. J. *Creativity and the individual*. Glencoe, Ill.: Free Press, 1960.

Steiner, G. A. *The people look at television*. New York: Knopf, 1963.

Steiner, I. D. Group dynamics. In P. R. Farnsworth, O. McNemar, & Q. McNemar (Eds.), *Annu. Rev. Psychol.*, 1964, 15, 421-446.

Steiner, I. D., & Field, W. L. Role assignment and interpersonal influence. J. abnorm. soc. Psychol., 1960, 61, 239-245.

Stewart, J. Q. Empirical mathematical rules concerning the distinction and equilibrium of population. Geograph. Rev., 1947, 37, 461-485.

Stoke, S. M., & West, E. D. Sex differences in conversational interests. J. soc. Psychol., 1931, 2, 120-126.

Stouffer, S. A. Problems in the application of correlation to sociology. J. Amer. Statist. Ass., 1934, 29, 52-58. (Reprinted in Stouffer, 1962, 264-270.)

Stouffer, S A., Lumsdaine, A. A., Lumsdaine, M.A., Williams, R., Smith, M., Janis, I., Star, S., & Cottrell, L. The American soldier: combat and its aftermath. Princeton: Princeton Univer. Press, 1949. Vol. 2.

Stouffer, S. A. Social research to test ideas. Glencoe, Ill.: Free Press, 1962. (Reprinted from P. F. Lazarsfeld, Radio and the printed page. New York: Duell, Sloan and Pearce, 1940. Pp. 266-272.)

Strodtbeck, F. L., & James, R. M. Social process in jury deliberations. Unpublished paper presented at annual meeting of American Sociological Society, 1955.

Strodtbeck, F. L., James, R. M., & Hawkins, C. Social status in jury deliberations. Amer. sociol. Rev., 1957, 22, 713-719.

Strodtbeck, F. L., & Mann, R. D. Sex role differentiation in jury deliberations. Sociometry, 1956, 19, 3-11.

Stuart, I. R. Minorities vs. minorities: cognitive, affective and conative components of Puerto Rican and Negro acceptance and rejection. J. soc. Psychol., 1963, 59, 93-99.

Sussman, L. Mass political letter writing in America. Publ. Opin. Quart., 1959, 23, 203-212.

Sussman, L. *Dear F. D. R.* New York: Bedminster, 1963.

Swift, A. L. Jr. The survey of the YMCA of the City of New York. New York: Association Press, 1927, limited edition.

Tannenbaum, P. H., & Noah, J. E. Sportugese: a study of sports page communication. *Journalism Quart.*, 1959, 36, 163-170.

Tarde, G. *L'Opinion et la Foule.* Paris: Felix Alcan, 1901.

Terman, L. M. The intelligence quotient of Francis Galton in childhood. *Amer. J. Psychol.*, 1917, 28, 209-215. (Abstracted in M. I. Stein, & S. J. Heinze, *Creativity and the individual*. Glencoe, Ill.: Free Press, 1960.)

Thomas, D. S. *Some new techniques for studying social behavior.* New York: Columbia Univer. Press, 1929.

Thomas, W. I., & Znaniecki, F. *The Polish peasant in Europe and America: monograph of an immigrant group.* Vol. I. Chicago: Univer. of Chicago, 1918.

Thorndike, E. L. *Your city.* New York: Harcourt Brace, 1939.

Thorndike, R.L. *Personnel selection.* New York: Wiley, 1949.

Toulouse, M. M., & Mourgue, R. Des réactions respiratoires au cours de projections cinématographiques. *Rev. Int. de Filmol.*, 1948, 2, 77-83.

Trueswell, R. W. A survey of library users' needs and behavior as related to the application of data processing and computer technique. Unpublished doctoral dissertation, Northwestern Univer., 1963.

Turner, W. Dimensions of foreman performance: a factor analysis of criterions measures. *J. appl. Psychol.*, 1960, 44, 216-223.

Ulett, G. A., Heusler, A., Ives-Word, V., Word, Th., & Quick, R. Influence of chlordiozepoxide on drug-altered EEG patterns and behavior. *Med. Exp.*, 1961, 5, 386-390. (a)

Ulett, G. A., Heusler, A., Callahan, J. Objective measures in psycho-pharmacology (methodology). In E. Rothlin (Ed.), *Neuro-psychopharmacology* 1961, 2, 401-409. (b)

Ulmer, S. S. Quantitative processes: some practical and theoretical applications. In *Jurimetrics*, Hans W. Baade (Ed.). New York: Basic Books, 1963.

Underwood, B. J. *Psychological research*. New York: Appleton-Century-Crofts, 1957.

Vernon, D. T. A., & Brown, J. The utilization of secondary or less preferred sources of information by persons in potentially stressful situations. Unpublished paper, 1963.

Vincent, C. E. Socioeconomic status and familial variables in mail questionnaire responses. *American J. Sociol.*, 1964, 69, 647-653.

Vose, C. E. *Caucasions only*. Berkeley, Calif.: Univer of Calif. Press, 1959.

Warner, W. L. *The living and the dead*. New Haven: Yale Univer. Press, 1959.

Warner, W. L., Meeker, M., Eells, K. *Social class in America*. Chicago: Science Res. Ass., 1949.

Washburne, C. The good and bad in Russian education. *New Era*, 1928, 9, 8-12.

Watson, J., Breed, W., & Posman, H. A study in urban conversation: sample of 1001 remarks overheard in Manhattan. *J. soc. Psychol.*, 1948, 28, 121-123.

Watson, R. I. Historical review of objective personality testing: the search for objectivity. In B. M. Bass, & I. A. Berg (Eds.), *Objective approaches to personality assessment*. Princeton, N.J.: D. van Nostrand, 1959.

Wax, R. H. Reciprocity in field work. In R. N. Adams, & J. J. Preiss (Eds.), *Human organization research*. Homewood, Ill.: Dorsey Press, 1960.

Webb, E. J. *Mens clothing study*. Chicago: Chicago Tribune, 1957.

Webb, E. J. The case for the effectiveness index. *J. adv. Res.*, 1962.(a)

Webb, E. J. How to tell a columnist, I, *Columbia J. Rev.*, 1962, 1, 23-25. (b)

Webb, E.J. Television programming and the effect of ratings. Paper read before Association for Education in Journalism, Chapel Hill, N.C., 1962. (c)

Webb, E. J. How to tell a columnist, II, *Columbia J. Rev.*, 1963, 2, 20.

Webb, E. J. The orthographies of seven African languages. In preparation.

Wechsler, H. Community growth, depressive disorders and suicide. *Amer. J. Sociol.*, 1961, 67, 9-16.

Weir, R. H. *Language in the crib*. The Hague: Mouton, 1963.

Weiss, D. J., & Dawis, R. V. An objective validation of factual interview data. *J. appl. Psychol* 1960, 44, 381-385.

Weitz, J. Selecting supervisors with peer ratings. *Personnel Psychol.*, 1958, 11, 25-35.

Werner, H., & Wapner, S. Changes in psychological distance under conditions of danger. *J. Pers.*, 1953, 24, 153-167.

Whisler, T. L., Harper, S. F. *Performance appraisal: research and practice*. New York: Holt, Rinehart & Winston, 1962.

Whyte, W. H. *The organization man*. New York: Schuster, 1951.

Wigmore, J. H. *The principles of judicial proof as given by logic, psychology, and general experience and illustrated in judicial trials*. Boston: Little, Brown, 1913.

Wigmore, J. H. A student's textbook of the law of evidence. Brooklyn: The Foundation Press, 1935.

Williams, R. Probability sampling in the field: a case history. *Publ*. *Opin*. *Quart*., 1950, 14, 316-330.

Wilson, E. G. *An* *introduction* *to* *scientific* *research*. New York: McGraw-Hill, 1952.

Windle, C. Test-retest effect on personality questionnaire. *Educ*. *psychol*. *measmt*, 1954, 14, 617-633.

Winship, E. C., & Allport, G. W. Do rosy headlines sell newspapers? *Publ*. *Opin*. *Quart*., 1943, 7, 205-210.

Winston, S. Birth control and the sex ratio at birth. *Amer*. *J*. *Sociol*., 1932, 38, 225-231.

Wolff, C. *A* *psychology* *of* *gesture*. London: Methuen, 1948.

Wolff, C. *The* *hand* *in* *psychological* *diagnosis*. London: Methuen, 1951.

Wolff, W., & Precker, J. A. Expressive movement and the methods of experimental depth psychology. In H. H. Anderson, & G. L. Anderson (Eds.), *An* *introduction* *to* *projection* *techniques*. New York: Prentice Hall, 1951.

Wolfson, R. Graphology. In H. H. Anderson, & G. L. Anderson (Eds.), *An* *introduction* *to* *projection* *techniques*. New York: Prentice Hall, 1951.

Zeisel, H. *Die* *Arbeitslosen* *von* *Marienbad*. Leipzig: Hirzel, 1932.

Zeisel, H. *Say* *it* *with* *figures*. (4th ed.) New York: Harper, 1957.

Zipf, G. K. Some determinants of the circulation of information. *Amer*. *J*. *Psychol*., 1946, 59, 401-421.

Zipf, G. K. *Human* *behavior* *and* *the* *principle* *of* *least* *effort*. Cambridge: Addison-Wesley, 1949.

Seating Aggregation as an Index of Attitude[1]

Donald T. Campbell and William F. Wallace

Northwestern University

The fact that all measures contain systematic irrelevancies, the fact that our usual verbal voluntary self-report measures are subject to voluntary distortion as well as unconscious self deception and are often inappropriate to administer, and the general truism that multiple methods are required to provide validity, all point to the need to explore novel measures. This study reports such an exploration, an effort to use voluntary classroom seating patterns to measure the general level of attitudes between Negroes and whites.

Where seating in a classroom is voluntary, the degree to which the Negroes and whites present sit by themselves versus mixing randomly may be taken as a presumptive index of the degree to which acquaintance, friendship, and preference are strongly colored by race as opposed to being distributed without regard to racial considerations. Such voluntary clustering by race will be termed aggregation, as opposed to the enforced separation connoted by the term segregation.

Just as an attitude questionnaire has its irrelevant components of response set, social desirability factor, social class differences in willingness to use hostile vocabulary for any purpose, etc., so such a measure as seating aggregation may be expected to have systematic sources of variance irrelevant to the concept of interest. Observation of the high degree of racial clustering or aggregation at the typical annual banquet of a society dedicated to removing racial barriers reflects no doubt the biased opportunities for prior acquaintance

---

[1] The preparation of this paper was supported by Project C998 Contract 3-20-001 with the Office of Education, U.S. Department of Health, Education and Welfare, under the provisions of Title VII of the National Defense Education Act.

rather than the racial attitudes of the persons involved. So too in a classroom, the tendency to sit with friends from one's neighborhood and previous schools provides a bias perhaps more directly reflecting acquaintance opportunity than lack of good will. None the less, if within a classroom there are marked shifts in aggregation from time to time, these might be interpreted as reflecting shifts in interracial fear and good will superimposed on the baseline provided by prior acquaintance opportunity. Or if two schools drawing from the same community provide markedly different aggregation indices, these differences might be attributed to attitudinal factors. The present study is of this latter case.

The Aggregation Index. What is needed is a way of estimating what a random seating arrangement would be and its sampling variance to provide an index of the degree of departure from randomness. Professor William Kruskal has provided such an index, as presented in the appendix to this paper. This index centers upon comparing the number of observed Negro-white seating adjacencies with an expected value computed from consideration of the number of Negroes, whites, clusters and isolates. Figure 1 shows an actual seating chart from the study, used in the following illustration of the index. The formula for the index is $I = \frac{(O-E)}{\sqrt{\sigma_E^2}}$

- - - - - - - - - - - - - - - - - -

Insert Fig. 1 about here

- - - - - - - - - - - - - - - - - -

where O = observed adjacencies, E = expected adjacencies and σ_E^2 = variance of the expected adjacencies, derived from consideration of the number of Negro students in class (M), the total number of students in class (N), the number of groups or clusters of students sitting together in a given row with no intervening empty seats (K), and the number of isolates or students sitting alone (K_1).

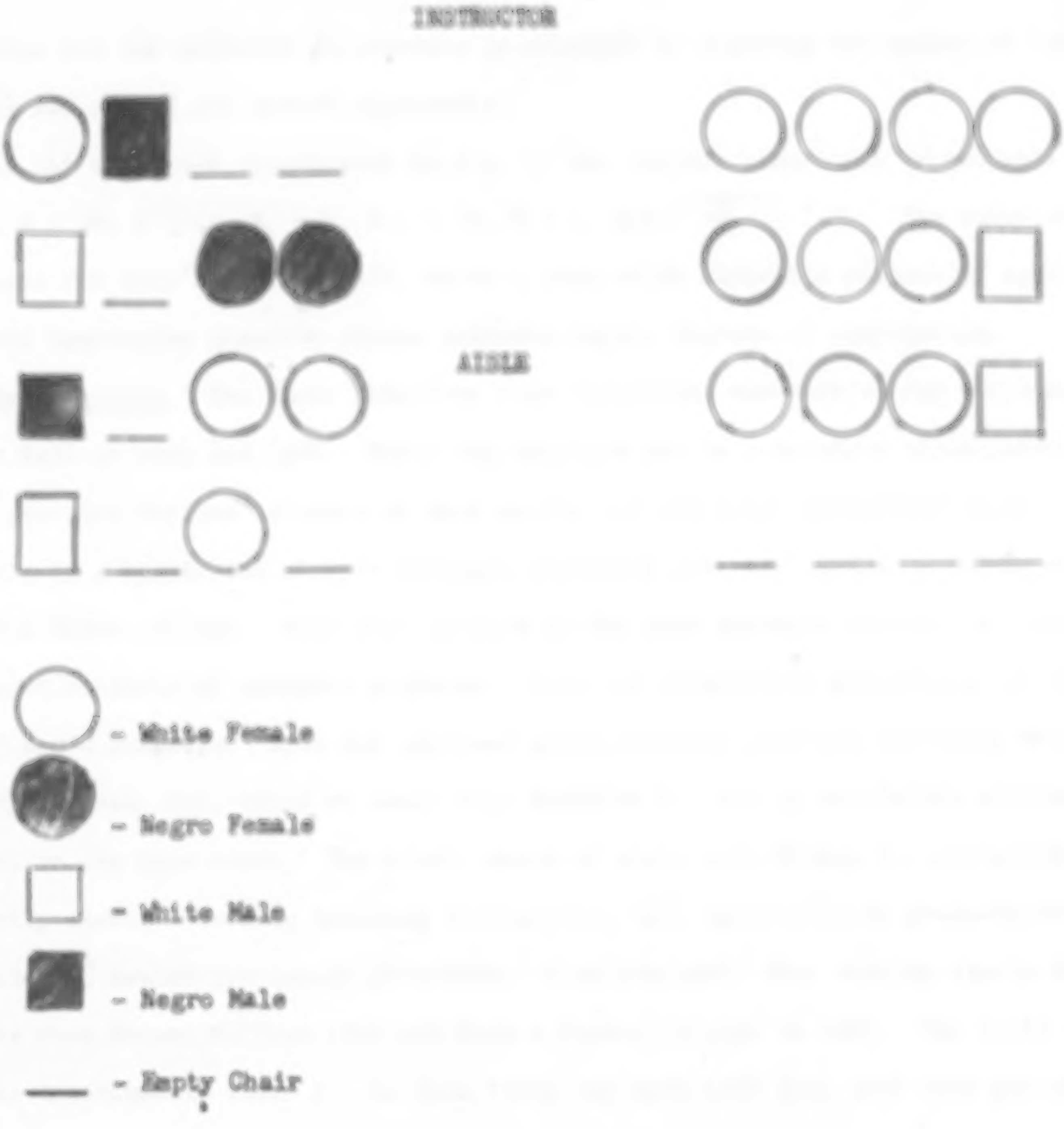

INSTRUCTOR
AISLE
- White Female
- Negro Female
- White Male
- Negro Male
- Empty Chair

Fig. 1. Sample seating chart from Normal U., 1964.

The formula for the expected adjacencies is $E = 2\frac{M(N-M)}{N(N-1)}(N-K)$. The variance is given by the following expression.

$$\sigma_E^2 = 2\frac{M(N-M)}{N(N-1)}(2N-3K+K_1) + 4\frac{M(M-1)(N-M)(N-M-1)}{N(N-1)(N-2)(N-3)}[(N-K)(N-K-1) - 2(N-2K+K_1)] - 4\frac{M^2(N-M)^2}{N^2(N-1)^2}$$

The value for the observed adjacencies is obtained by counting the number of times Negroes and whites are seated adjacently.

In the classroom illustrated in Fig. 1, the various values are as follows: $M = 6$, $N = 22$, $K = 10$, $K_1 = 4$, $E = 4.99$, $O = 1$, and $\sqrt{\sigma_E^2} = 1.51$. The value of the index for this class is -2.64, where a zero value indicates absence of aggregation and increasing negative values indicate larger degrees of aggregation.

Data Sources. The basic data come from classrooms observed at two colleges in the Fall of 1963 and 1964. These two colleges had on a previous unpublished study provided the two extremes of most prejudiced and least prejudiced white students in a comparison of five colleges scattered over the nation including one southern white college. Both were located in the same northern metropolis. Both consisted entirely of commuter students. Both had substantial minorities--as colleges go--of Negro students. Both had unbiased administrative policies including Negro faculty members. One, which we shall call Downtown U., was by reputation militantly liberal on the race issue. The other, which we shall call Normal U., attracted primarily students seeking teaching credentials, and, apparently of predominantly traditional, nonintegrationist attitudes. Also available were seating charts for classes from Normal U. from 1951 and from a Junior College in 1963. The basic data pool is described in Table 1. In this table the 1963-1964 data have been pooled.

- - - - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - - - -

Table 1. Basic Data Pool

	Number of Classes	Number of Negroes	Number of Whites	Total Number of Students
Downtown U. 1963-1964	23	58	495	553
Normal U. 1963-1964	20	152	379	531
Normal U. 1951	19	89	435	524
Junior College 1963	12	138	129	267

Inspection showed no trends within this period.

The 1963-1964 data were collected by an observer who entered the classroom as though a student, and working at the back of the room, prepared the seating chart. Classes having no Negroes or no whites were excluded. The 1951 observations were made by the administrator of a test while the test was being taken by the students and did not include a record of sex. The courses sampled covered a variety of liberal arts lecture courses. Observations at both schools were made within a week of each other in both 1963 and 1964, with no incidents of racial relevance occurring during the period of observation.

Results

The Generality of Aggregation by Race. Table 2 shows the basic results for the class indices. All schools have a negative mean index, indicating fewer Negro-white seating adjacencies than would be expected by chance. Using a t test based upon an N of groups, all of these means are significantly different from zero bias at the $p < .01$ level. Most individual classes show negative indices, although there are a few that do not. No doubt much of this common aggregation reflects the out-of-school bias in acquaintanceship opportunities.

- - - - - - - - - - - - - - - - -
Insert Table 2 about here
- - - - - - - - - - - - - - - - -

School Differences in Aggregation. The two schools show a significant difference in aggregation indices in the expected direction. The Downtown U. mean of -.81 was significantly less negative than the Normal U. mean of -1.50 with a t of 2.46, 39 degrees of freedom, $p < .02$. The difference was in the same direction and to about the same degree in the 1963 and 1964 data taken separately. The utility of the index as a measure of attitude thus receives some confirmation. The degree of differentiation is, however, by no means as sharp as that provided by the earlier anonymous questionnaire data. The degree of overlap between the schools is great--as indicated by the occurrence in each of classrooms with a positive index.

Table 2. Racial Aggregation Indices

	Downtown U. 1963-1964	Normal U. 1963-1964	Normal U. 1951	Junior College 1963
Mean Index	-.81	-1.50	-1.05	-1.41
Standard Deviation	.81	.98	1.53	1.46
t from zero index	4.50	6.82	3.00	3.20
Number of classes	21	20	19	11
Number of classes with positive indices	3	1	6	2

Aggregation Changes Over Time. For Normal U. the difference between the 1951 data and that for 1964 was not significant. Corresponding attitude test data from the school for the time trend is not available, although general public opinion surveys show a decrease in prejudice for this decade.[2] The direction of shift for the index is indeed counter to this national trend.

Aggregation by Sex. As is commonly noted, classroom seating shows aggregation by sex. The computation of indices for sex aggregation was thought to be of value both as a check on the index and to provide a comparison base for the racial aggregation indices. Table 3 shows the basic values for these indices. (Sex was not recorded in the 1951 data from Normal U., and classes have been omitted which did not have members of both sexes.) The general finding is one of significant sex aggregation in all schools. The difference between sex and race aggregation is not significant when computed by direct difference t on the 51 classes for which both indices were available (Race mean = -1.20, Sex mean = -.99, t = 1.24, df = 49, $< .25$).

- - - - - - - - - - - - - - - - - - -
Insert Table 3 about here
- - - - - - - - - - - - - - - - - - -

Interaction of Sex and Race. This analysis asks whether or not the number of adjacencies is particularly low when the adjacency crosses both race and sex, That is, do adjacencies between Negro men and white women or between Negro women and white men occur still less frequently than do adjacencies between persons differing on race alone, or on sex alone. No entirely adequate approach to this problem was

---

[2] Hyman, H. H. and Sheatsly, P. B. Attitudes on desegregation. Scientific American, July, 1964, volume 211, No. 1, pp. 16-23.

11

Table 3. Aggregation Indices for Sex

	Downtown U. 1963-1964	Normal U. 1963-1964	Junior College 1963
Mean Index	-.90	-1.36	-.77
Standard Deviation	1.36	.98	1.18
t from zero index	3.10	6.18	2.26
Number of Classes	22	20	12
Number of Classes with positive indices	4	2	4

found. What has been explored are fragmented indices, paying attention only to the two subgroups involved, and treating all other persons as though they were empty seats. This analysis has been limited to those classes providing persons in all four race-sex subgroups. Table 4 shows the results of this analysis. While the two indices crossing both race and sex average higher than either of the other pairs, they both fall within the range of indices provided by the other subgroups. Using class-by-class direct difference t's for the averages of these pairs, no significant differences were found (Cross Sex & Race vs cross Race $t = 1.85$; Cross Sex & Race vs cross Sex, $t = 1.53$; Cross Sex vs cross Race, $t = .20$, df = 16 in each case). Thus while a slight trend toward an interaction was present, it did not approach usual standards for statistical significance.

- - - - - - - - - - - - - - - -

Insert Table 4 about here

- - - - - - - - - - - - - - - -

Control Analyses. It is very difficult to achieve descriptive indices of degree of effect which are comparable with assurance across heterogeneous conditions of size and proportion. For this reason, certain control analyses have been done. It is possible that the index might not be comparable between two classes varying widely in the proportion of Negroes present. Computing the correlation between this proportion and the resulting index, a trivial correlation of -.15 (df = 70) was obtained, effectively removing this as an obstacle to interpretation. Class size is somewhat more problematic. It correlates with the index -.29 (df = 70), significant at the $p < .02$ level. The larger the class the more aggregation by race. However, when comparisons are restricted to the 1963-1964 Downtown U. and Normal U. samples, the class size-aggregation index correlation does not reach an acceptable level of significance ($r = -.15$, df = 40). Also, the class size correlation with aggregation by sex did not approach significance ($r = -.08$, df = 53). These latter results do not support an interpretation that the aggregation index is artificially

Table 4. Aggregation Indices with Fragmented Classes

	Cross Race & Sex		Cross Race		Cross Sex	
	MN-FW	MW-FN	MN-MW	FN-FW	MN-FN	MW-FW
Mean Index	-1.20	-1.26	-.66	-1.21	-.60	-1.34
Average Subclass Size	15.00	17.59	14.82	17.76	9.47	23.12
Number of Classes	17	17	17	17	17	17
Number of classes with positive index	1	4	6	3	5	2

influenced by class size. In any event, it does not affect the interpretation of the Downtown U.-Normal U. aggregation differences, as the 1963-1964 mean sizes for these two are 26.33 and 26.55 respectively.

In some respects the index neglects features that might artificially affect the number of Negro-white adjacencies. Thus no use is made of the number of vacant seats nor their possible location so as to buffer Negro-white adjacency. Undoubtedly in many instances a shortage of seats increases Negro-white adjacencies in a manner not reflecting the preferences of the persons involved. As a partial check on this, the proportion of empty seats was correlated with the race aggregation index. The resulting correlation was -.23, which with an N of 52 almost reaches the $p < .05$ level of significance ($r = .27$ for $p = .05$). The correlation, however, is in the opposite direction from that expected; the larger the proportion of empty seats the less aggregation. For sex aggregation, the correlation was but +.02. Thus neglect of empty chairs does not in these data appear to have produced misleading effects. Another feature which an ideal index might attend to is the number or proportion of row-ends where no adjacency is possible. Thus an arrangement of the same number of chairs into a pattern of many short rows reduces the opportunities for adjacency in comparison with an arrangement of fewer longer rows. This attribute varied so little in the present sample that no control correlation has been computed.

Discussion

While there is no indication of an index of high precision or purity, the results do encourage further exploration. One of the most needed studies is the examination of shifts in the index over time in an area in which local crises might be expected to provide shifts in the degree of mutual trust. A cooperating group of instructors could easily produce a daily series of seating charts in classes in which written classroom work takes place. The index can also be applied in any other situation in which unidimensional adjacency can be noted. One recurrent class of situations of this type is in the waiting lines for cafeterias, theater tickets, time-clock punching, etc. For waiting lines, the formula is simpler, as K becomes 1.

The adjacency index is but one of a larger class of potential aggregation indices. Consider a school playground during lunch time in which many of the pupils can be identified as participating in activity groups, be they baseball games, hopscotch, penny-pitching, etc. Utilizing children so classified, one could test the tendency of these to have racially biased composition by a chi-square or contingency coefficient based upon a contingency table in which columns were Negro and white, and in which each play group constituted a row. In a similar way, ethnic bias in isolates versus group-involved could be computed. A time-series of such indices from multiracial schoolyards might provide an index of changing tensions.

Another way of conceptualizing aggregation is in terms of the steepness of the interracial boundary--and note that the schoolroom and school yard indices must come from essentially boundary conditions--they are not available for either all white or all Negro schools. If one considers for example the residences, sidewalks, stores and public services on a street which crosses the Negro-white boundary at right angles, the more aggregation there is, the "steeper" the racial boundary, i.e., the narrower the transition area in which there is mixed racial presence, the more nearly the shift is from 100% Negro to 100% white, without blocks, stores, sidewalk segments, etc. which have both Negroes and whites present. For presences for which daily decisions and changes are possible, such as shopping for groceries, an increase in the steepness of the boundary might indicate an increase in tension, as Negroes feel reluctance to shop in predominantly white stores and as whites feel reluctance to shop in predominantly Negro ones. Such considerations move us close to the sociological literature on city indices of segregation, computed from census data.

Summary

Bias in the number of Negro-white seating adjacencies has been employed as a tentative index of interracial attitudes. In all four schools studied, significant aggregation by race was found. Two schools selected on a priori grounds showed the expected differences in degree of aggregation. All schools also showed significant aggregation by sex. Control analyses examining the effect of class size, proportion of Negroes and proportion of vacant seats do not indicate artifacts in these respects. The index of seating aggregation is judged worthy of further exploration.

Appendix

A Combinatorial Problem Suggested by a Study in Race Relationships[1]

Willaim H. Kruskal

University of Chicago

A study in Negro-white relationships led to a combinatorial problem whose solution may be of interest to workers on allied questions. The problem is to find the expected value and variance of the number of adjacencies between Negro and white students who have seated themselves freely in a classroom. Formulas for expected value and variance are derived in this note for a conditional distribution to be shortly described.

An observer diagrams the seating arrangement in the classroom so that each seat is counted either as Negro, white, or empty. The number of times that a Negro and white appear immediately next to each other in the same row is called the number of adjacencies, A, and is taken to be a measure of Negro-white relationships in the school. The question arises: what is the expected value and variance of A under the hypothesis of random seating? In this note we shall restrict ourselves to the conditional distribution of random seating but with specifically named seats always occupied; i.e., random permutations of the students within the same groups of seats.

We denote by K the number of unbroken sequences of row-wise contiguous occupied seats, the sequences containing $n_1, \ldots, n_K$ seats respectively.

---

[1] This paper was prepared in 1951 in response to a request from Donald T. Campbell in connection with research sponsored by the Committee on Education, Training and Research in Race Relations of the University of Chicago. Professor Kruskal's efforts were supported by a grant from the Office of Naval Research to the Department of Statistics, University of Chicago.

$N = \sum_{i=1}^{K} n_i$. Our random variables, arranged in these K groups are:

$$X_1, X_2, \cdots, X_{n_1}$$

$$X_{n_1+1}, X_{n_1+2}, \cdots\cdots, X_{n_1+n_2}$$

$$\vdots$$

$$X_{N-n_K+1}, X_{N-n_K+2}, \cdots\cdot, X_N$$

Each X can take two values: 1 for Negro, and 2 for white. Putting M for the number of Negroes and N-M for the number of whites, the joint distribution of the X's under the hypothesis of (conditional) random seating is that any of the $\binom{N}{M}$ combinations of assigning M 1's and N-M 2's to the above array of X's is equally probable.

The statistic of interest, the number of adjacencies, A is:

$$\{|X_1-X_2|+|X_2-X_3|+\ldots+|X_{n_1-1}-X_{n_1}|\} + \{|X_{n_1+1}-X_{n_1+2}|+\cdots+|X_{n_2-1}-X_{n_1+n_2}|\}$$
$$+\cdots\cdot+ \{|X_{N-n_K+1}-X_{N-n_K+2}|+\ldots+|X_{N-1}-X_N|\}$$

with the convention that if any group contains a *single* *member* ($n_i=1$), *no* absolute difference corresponding to it appears in the sum.

For convenience set

$$A = \{U_1+\cdots+U_{n_1-1}\} + \{U_{n_1+1}+\cdots+U_{n_1+n_2-1}\}$$
$$+\ldots+\{U_{N-n_K+1}+\cdots+U_{N-1}\}$$

where $U_j = |X_j-X_{j+1}|$.

We proceed to compute first and second moments.

$$EU_j = \text{Prob}\ (X_j \neq X_{j-1}) = 2\,\frac{M(N-M)}{N(N-1)}$$

Hence: $EA = 2\,\frac{M(N-M)}{N(N-1)}\,(\Sigma n_i - K) = 2\,\frac{M(N-M)}{N(N-1)}\,(N-K) = EA$

Let us set $\eta = E\,U_j^2$

$\zeta = E[U_j U_{j+1}]$

$\xi = E[U_j U_k] \quad , \quad |k-j| > 1$

We next compute EA^2 which must equal:

(No. of U's) $\cdot\, \eta$ + (No. of $U_j U_{j+1}$ cross-products)$\cdot\,\zeta$

+ (No. of $U_j U_k$ ($|j-k| > 1$) cross-products)$\cdot\,\xi$.

Clearly there are N-K U's. We next ask how many $U_j U_{j+1}$ cross-products appear.

Let K_1, K_2 = the numbers of sequences with just one and two X_j members respectively, and hence zero and one U_j members respectively. Let K_* = the number of groups with more than 2 X_j members. $K = K_1 + K_2 + K_*$.

Only the K_* sub-groups contribute to the $U_j U_{j+1}$ products, each giving rise to $2(n_i-2)$ of them. Let the symbol Σ_* stand for summation over only the K_* sequences. Hence the desired number is: $2\,\Sigma_*(n_i-2)$

$= 2[N-K_1-2K_2) - (2K_*)]$

$= 2[N \; -2K + K_1]$

Now how many products of form $U_j U_k$ ($|j-k| > 1$) are there? Each of the K_2 groups contributes $N - K - 1$. Each of the K_* groups contributes

$2(N-K-2) + (n_i-3)\,(N-K-3)$

So we have:

$K_2(N-K-1) + 2K_*(N-K-2) + (N-K-3)\,\Sigma_*\,(n_i-3)$

$= (N-K)\,(N-K-1) - 2N + 4K - 2K_1 = (N-K)(N-K-1) - 2[N-2K+K_1]$

This is a check, since the sum of the two numbers of products must be $(N-K)(N-K-1)$

It follows that:

$$EA^2 = (N-K)\eta + 2(N-2K+K_1)\zeta + [(N-K)(N-K-1) - 2(N-2K+K_1)]\xi$$

It remains only to work out the values of η, ζ, ξ.

η. We get EU_j^2 by explicitly writing down that part of the distribution of X_j, X_{j+1} for which $U_j = 1$

X_j	X_{j+1}	Probability
1	0	$\frac{M(N-M)}{N(N-1)}$
0	1	same

Hence $EU_j^2 = 2\frac{M(N-M)}{N(N-1)}$

ζ. We get EU_jU_{j+1} by explicitly writing down that part of the distribution of X_j, X_{j+1}, X_{j+2} for which $U_jU_{j+1} = 1$

X_j	X_{j+1}	X_{j+2}	Probability
1	0	1	$\frac{M(M-1)\ (N-M)}{N(N-1)\ (N-2)}$
0	1	0	$\frac{M(N-M)\ (N-M-1)}{N(N-1)\ (N-2)}$

Hence $EU_jU_{j+1} = \frac{M(N-M)}{N(N-1)(N-2)} \quad (M-1+N-M-1) = \frac{M(N-M)}{N(N-1)}$

ξ. We get EU_jU_k by explicitly writing down that part of the distribution of X_j, X_{j+1}, X_k, X_{k+1} for which $U_jU_k = 1$

X_j	X_{j+1}	X_k	X_{k+1}	Probability
1	0	1	0	$\frac{M(M-1)(N-M)(N-M-1)}{N(N-1)(N-2)(N-3)}$
1	0	0	1	"
0	1	1	0	"
0	1	0	1	"

Hence:

$$\xi = 4\,\frac{M(M-1)(N-M)(N-M-1)}{N(N-1)(N-2)(N-3)}$$

Whence:
$$\begin{aligned}\text{Var } A = {} & 2\,\frac{M(N-M)}{N(N-1)}\,(N-K) + 2(N-2K+K_1)\,\frac{M(N-M)}{N(N-1)} \\ & + 4[(N-K)(N-K-1)-2(N-2K+K_1)]\;\frac{M(M-1)(N-M)(N-M-1)}{N(N-1)(N-2)(N-3)} \\ & - 4\,\frac{M^2(N-M)^2}{N^2(N-1)^2}\;(N-K)^2\end{aligned}$$

or:
$$\begin{aligned}\text{Var } A = {} & 2\,\frac{M(N-M)}{N(N-1)}\,(2N-3K+K_1) \\ & + 4\,\frac{M(M-1)(N-M)(N-M-1)}{N(N-1)(N-2)(N-3)}\;[(N-K)(N-K-1)-2(N-2K+K_1)] \\ & - 4\,\frac{M^2(N-M)^2}{N^2(N-1)^2}\;(N-K)^2\end{aligned}$$

As a check we may consider the case $K = 1$, $K_1 = 0$, so that A is just the total number of runs in one sequence minus one.

Then an article by Wald and Wolfowitz[*], gives the well known:

$$EA = \frac{2M(N-M)}{N}$$

$$\text{Var } A = \frac{2M(N-M)[2M(N-M)-N]}{N^2(N-1)}$$

Our formulas give the same values after algebraic simplification.

Various approximations for computing purposes may be made depending on the relative magnitudes of M, N, N-M, N-K, and K_1. These are obvious and need not be spelled out in detail.

---

[*] "On a Test whether Two Samples Are from the Same Population," Annals of Mathematical Statistics, 11:147.

As N approaches infinity we may look for asymptotic unit-normality of:

$$\frac{A - EA}{\sqrt{\operatorname{Var} A}}$$

If we suppose that K is fixed but that M, n_1, n_2, ..., n_K are functions of N, and that $\frac{M}{N}$ has a limit, u, as N approaches infinity; then we must have asymptotic normality, since then A is just the total number of runs less a number which can vary only between 1 and K. In the sense of convergence in distribution:

$$\frac{A-2u(1-u)N}{2u(1-u)\sqrt{N}}$$

is asymptotically unit-normal.

A Procedure for Estimating the Amount of Trait, Method, and Error Variance Attributable to a Measure[1]

by

Leroy Wolins

Iowa State University

In the multimethod-multitrait situation the investigator is dealing with a number of traits, t, and a number of methods, m. Usually he measures his observational units using every combination of methods and traits so that he observes each observational unit m times t times. The model proposed requires that each of this m times t empirical variables be linear combinations of m plus t conceptual variables. That is, for each method-trait combination a persons true score on a trait is weighted a certain amount and a persons true score on a method is weighted a certain amount. The weighting is done by intractable nature and she throws into each empirical measure some error to further confuse matters. Formally,

$$Y_{i(jk)} = b_{(jk)j.k} X_{ij} + b_{(jk)k.j} X_{ik} + e_{i(jk)}.$$

This equation states that person i's observed score on trait j measured by method k, $Y_{i(jk)}$, is equal to the weighted sum of a trait score, X_{ij}, a method score, X_{ik}, and error. The weights are partial regression coefficients used by nature and are based on the correlations of the true scores

---

[1] The preparation of this report was supported in part by Project C-998, Contract 3-20-001, with the Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act. The contract is with Northwestern University, Donald T. Campbell principal investigator.

with the observed score and the correlation between the true scores. That is,

$$b_{(jk)j.k} = \frac{r_{(jk)j} - r_{(jk)k}\, r_{jk}}{1 - r^2_{jk}} \quad \text{and}$$

$$b_{(jk)k.j} = \frac{r_{(jk)k} - r_{(jk)j}\, r_{jk}}{1 - r^2_{jk}} .$$

The questions about each empirical measure which the model suggests are:

1) What proportion of the variance is attributable to trait $(b_{(jk)j.k}\, r_{(jk)j})$?

2) What proportion of the variance is attributable to method $(b_{(jk)k.j}\, r_{(jk)k})$?

3) What proportion of variance is attributable to error $(1 - b_{(jk)j.k}\, r_{(jk)j} - b_{(jk)k.j}\, r_{(jk)k})$?

In order to clarify the situation, the scores and correlations, both empirical and hypothetical, are indicated on the following page for a multi-method-multitrait situation involving three methods and three traits.

From these "data" we may generate a correlation matrix which would appear as on page 4.

Inspection of this correlation matrix may immediately suggest a solution. That is, the correlations between the observed and conceptual variables are, indeed, oblique factor loadings and the intercorrelations between

Observational Units	Observed scores									Conceptual scores					
	A1	B1	C1	A2	B2	C2	A3	B3	C3	A	B	C	1	2	3
1	$Y_{1(A1)}$	$Y_{1(B1)}$	$Y_{1(C1)}$	$Y_{1(A2)}$	$Y_{1(B2)}$	$Y_{1(C2)}$	$Y_{1(A3)}$	$Y_{1(B3)}$	$Y_{1(C3)}$	X_{1A}	X_{1B}	X_{1C}	X_{11}	X_{12}	X_{13}
2	$Y_{2(A1)}$	$Y_{2(B1)}$	$Y_{2(C1)}$	$Y_{2(A2)}$	$Y_{2(B2)}$	$Y_{2(C2)}$	$Y_{2(A3)}$	$Y_{2(B3)}$	$Y_{2(C3)}$	X_{2A}	X_{2B}	X_{2C}	X_{21}	X_{22}	X_{23}
3	$Y_{3(A1)}$	$Y_{3(B1)}$	$Y_{3(C1)}$	$Y_{3(A2)}$	$Y_{3(B2)}$	$Y_{3(C2)}$	$Y_{3(A3)}$	$Y_{3(B3)}$	$Y_{3(C3)}$	X_{3A}	X_{3B}	X_{3C}	X_{31}	X_{32}	X_{33}
$\vdots$															
N	$Y_{N(A1)}$	$Y_{N(B1)}$	$Y_{N(C1)}$	$Y_{N(A2)}$	$Y_{N(B2)}$	$Y_{N(C2)}$	$Y_{N(A3)}$	$Y_{N(B3)}$	$Y_{N(C3)}$	X_{NA}	X_{NB}	X_{NC}	X_{N1}	X_{N2}	X_{N3}

Intercorrelations

	Observed Variables									Conceptual Variables					
–	A1	B1	C1	A2	B2	C2	A3	B3	C3	A	B	C	1	2	3
A1		$r_{(A1)(B1)}$	$r_{(A1)(C1)}$	$r_{(A1)(A2)}$	$r_{(A1)(B2)}$	$r_{(A1)(C2)}$	$r_{(A1)(A3)}$	$r_{(A1)(B3)}$	$r_{(A1)(C3)}$	$r_{(A1)A}$	$r_{(A1)B}$	$r_{(A1)C}$	$r_{(A1)1}$	$r_{(A1)2}$	$r_{(A1)3}$
B1			$r_{(B1)(C1)}$												
C1															
A2															
B2															
C2								(etc.)							
A3															
B3															
C3															
–															
A										1.00	r_{AB}	r_{AC}	r_{A1}	r_{A2}	r_{A3}
B											1.00	r_{BC}	r_{B1}	r_{B2}	r_{B3}
C												1.00	r_{C1}	r_{C2}	r_{C3}
1													1.00	r_{12}	r_{13}
2														1.00	r_{23}
3															1.00

the conceptual variables are the correlations among the oblique factors. Also, the rotation problem is not insurmountable since we know what variables should load highest on what factors. However, in general, one cannot obtain a unique solution for six factors from nine variables. In general, for one to locate uniquely a set of points in a factor space, there must be no rotation such that less than three variables load substantially on any given factor. Thus, if data resulting from measuring six traits by six methods were available one could hope to extract 12 factors uniquely and answer the questions the model suggests.

In order to derive a procedure from the model that is more generally applicable some assumption in addition to the usual ones are required. One usual assumption is the error has mean zero and is uncorrelated with the variates. We might also assume something about the distribution of the error and the variates but the writer aspires no further than presenting the population model at the present time so that assumption about the distributions are not necessary. Given the independence and zero mean assumption regarding error, one can write down the known values in terms of the unknown values as follows:

1) For the heterotrait-heteromethod values

$$r_{(jk)(j'k')} = b_{(jk)j.k}b_{(j'k')j'.k'}r_{jj'} + b_{(jk)j.k}b_{(j'k')k'.j}r_{jk'} + b_{(jk)k.j}b_{(j'k')j'.k'}r_{kj'} + b_{(jk)k.j}b_{(j'k')k'.j'}r_{kk'}$$

2) For the heterotrait-homomethod values

$$r_{(jk)(j'k)} = b_{(jk)j.k}b_{(j'k)j'.k}r_{jj'} + b_{(jk)j.k}b_{(j'k)k.j'}r_{jk} + b_{(jk)k.j}b_{(j'k)j'.k}r_{kj'} + b_{(jk)k.j}b_{(j'k)k.j'} \qquad (1.00)$$

3) For the homotrait-heteromethod values

$$r_{(jk)(jk')} = b_{(jk)j.k}b_{(jk')j.k'}(1.00) + b_{(jk)j.k}b_{(jk')k'.j}r_{jk'} + b_{(jk)k.j}b_{(jk')j.k'}r_{kj} + b_{(jk)k.j}b_{(jk')k'.j}r_{kk} .$$

No procedure for estimating these unknown values occur to the writer except the one previously mentioned which is applicable when a large number of empirical variables are used. Thus two additional assumptions are investigated:

1) The methods do not correlate with traits. That is,

$$r_{jk} = r_{j'k} = r_{jk'} = 0 \quad \text{and}$$

2) The methods are uncorrelated. That is

$$r_{kk'} = 0 \quad \text{but} \quad r_{kk} = r_{k'k'} = 1.00.$$

Taken together these two assumptions are like the assumption made about the error associated with the empirical measures. That is, method variance is like error variance in that all it correlates with is itself. Unlike error variance, however, the same method variance is weighted into different empirical measures, thus there is a method bias.

Making these two assumptions and again expressing the known values in terms of the unknown values we obtain the following:

1) For the heterotrait-heteromethod values

$$r_{(jk)(j'k')} = b_{(jk)j.k}b_{(j'k')j'.k'}r_{jj'}$$

2) For the heterotrait-homomethod values

$$r_{(jk)(j'k)} = b_{(jk)j.k}b_{(j'k)j'.k}r_{jj'} + b_{(jk)k.j}b_{(j'k)k.j'}(1.00)$$

3) For the homotrait-heteromethod values

$$r_{(jk)(j.k')} = b_{(jk)j.k}\, b_{(jk')j.k'}\ (1.00).$$

It is evident from inspection of multimethod-multitrait matrices reported in the literature that this model does not fit a large amount of data. For example, using some of Kelley and Fiske data on assessment of clinical psychologists, the following correlations are reported:

Method 3: Self Ratings
Method 1: Staff Ratings

		A_1	B_1	E_1
Assertive	A_3	-	.31	.12
Cheerful	B_3	.17	-	-.03
Broad Interests	E_3	.37	.15	-

Theoretically,

$$\begin{array}{cccc}
 & A_1 & B_1 & C_1 \\
A_3 & - & b_{(A3)A.3}\, b_{(B1)B.1}\, r_{AB} & b_{(A3)A.3}\, b_{(C1)C.1}\, r_{AC} \\
B_3 & b_{(B3)B.3}\, b_{(A1)A.1}\, r_{AB} & - & b_{(B3)B.3}\, b_{(C1)C.1}\, r_{BC} \\
C_3 & b_{(C3)C.3}\, b_{(A1)A.1}\, r_{AC} & b_{(C3)C.3}\, b_{(B1)B.1}\, r_{BC} & -
\end{array}$$

and

$$\frac{r_{(A3)(B1)}\ r_{(B3)(C1)}}{r_{(A3)(C1)}} = \frac{r_{(B3)(A1)}\ r_{(C3)(B1)}}{r_{(C3)(A1)}} = \frac{b_{(B3)B.3}\ b_{(B1)B.1}\ r_{(AB}\ r_{BC}}{r_{AC}}.$$

Obviously, in this case, this is not so:

$$\frac{(.31)(-.03)}{.12} \neq \frac{(.17)(.15)}{.37}.$$

A possible explanation for this is that the methods do intercorrelate and are weighted differently from measure to measure. That is $r_{13} < 0$ and the method variance is weighted highly when the staff is rating "Assertiveness" and when the ratee is rating "Broad Interests"; thus $r_{(A1)(E3)} = .37$. On the other hand the method variance is weighted lowly when the staff is rating "Broad Interests" and the ratee is rating "Assertiveness"; thus $r_{(A3)(E1)} = .12$. Of course, another possible explanation is that "Broad Interests" correlates with "Self Ratings" but not with "Staff Ratings," and neither method correlates with "Assertiveness." However, we will assume, for the present, that nature is not so pernicious as to cause traits to correlate with methods but apparently the assumption that methods are independent must be relaxed. This results in the following equations:

1. For the heterotrait-heteromethod values

$$r_{(jk)(j'k')} = b_{(jk)j.k}\, b_{(j'k')j'.k'}\, r_{jj'} + b_{(jk)k.j}\, b_{(j'k')k'.j'}\, r_{kk'}$$

2. For the heterotrait-homomethod values

$$r_{(jk)(j'k)} = b_{(jk)j.k}\, b_{(j'k)j'.k}\, r_{jj'} + b_{(jk)k.j}\, b_{(j'k)k.j'}\, (1.00)$$

3. For the homotrait-heteromethod values

$$r_{(jk)(j.k')} = b_{(jk)j.k}\, b_{(jk')j.k'}\, (1.00) + b_{(jk)k.j}\, b_{(jk')k'.j}\, r_{kk'}$$

These equations still do not suggest a direct solution. However, an additional condition suggests an analytical procedure. That is, if the $r_{jj'}$ correlations may be explained by one factor and the $r_{kk'}$ correlations may be explained by one factor then the heteromethod-heterotrait values,

$$r_{(jk)(j'k')} = (a_j\, b_{(jk)j.k})(a_{j'}\, b_{(j'k')j'.k'}) + (a_k\, b_{(jk)k.j})(a_{k'}\, b_{(j'k')k'.j})$$

$$= T_{(jk)}\, T_{(j'k')} + M_{(jk)}\, M_{(j'k')}\ .$$

If there are only three methods and three traits this condition is not so much an assumption as a computational gimmick since one factor will reduce the residuals to zero. Even if the Heywood case occurs we may, in some sense, estimate $T_{(jk)}\, T_{(j'k')}$ independently of $M_{(jk)}\, M_{(j'k')}$. Computationally this involves extracting two factors from a correlation matrix whose dimension is determined by at least three methods and three traits, weighting all residuals zero except the residuals from the heteromethod-heterotrait values. Formally, with two factors we wish to minimize.

$$\sum_{jk=11}^{mt} \sum_{j'k'=11}^{mt} \frac{(r_{(jk)(j'k')} - \hat{r}_{(jk)(j'k')})^2}{(1 - h^2_{jk})^2\,(1 - h^2_{j'k'})^2}$$

where $j \neq j'$ and $k \neq k'$.

Next, in order to estimate $b_{(jk)j.k}\, b_{(j'k')j'.k}\, r_{jj'}$ independently of $b_{(jk)k.j}\, b_{(j'k')k'.j}\, r_{kk'}$ the two factors must be rotated properly. In order to communicate the basis for rotation the results for three methods and three traits rotated properly are presented below:

	I	II
A1	$b_{(A1)A.1}\, a_A$	$b_{(A1)1.A}\, a_1$
A2	$b_{(A2)A.2}\, a_A$	$b_{(A2)2.A}\, a_2$
A3	$b_{(A3)A.3}\, a_A$	$b_{(A3)3.A}\, a_3$
B1	$b_{(B1)B.1}\, a_B$	$b_{(B1)1.B}\, a_1$
B2	$b_{(B2)B.2}\, a_B$	$b_{(B2)2.B}\, a_2$
B3	$b_{(B3)B.3}\, a_B$	$b_{(B3)3.B}\, a_3$

	I	II
C1	$b_{(C1)C.1}\ a_C$	$b_{(C1)1.C}\ a_1$
C2	$b_{(C2)C.2}\ a_C$	$b_{(C2)2.C}\ a_2$
C3	$b_{(C3)C.3}\ a_C$	$b_{(C3)3.C}\ a_3$

Next, attending to the residual matrix the only non-zero entries are the heterotrait-homomethod residuals and the homotrait-heteromethod residuals. These are of the form

$$\hat{r}_{(jk)(j'k)} = b_{(jk)k.j}\ b_{(jk)k.j}\ (1 - a_k^2)$$

or

$$\hat{r}_{(jk)(jk')} = b_{(jk)j.k}\ b_{(jk')j.k'}\ (1 - a_j^2) .$$

By factoring these six matrices, each separately, we can compute the communalities which are of the form

$$\hat{h}^2_{(jk)j} = b^2_{(jk)j.k}\ (1 - a_j^2)$$

or

$$\hat{h}^2_{(jk)k} = b^2_{(jk)k.j}\ (1 - a_k^2) .$$

These communalities are directly proportional to the corresponding entries squared from the two factor results properly rotated:

$$\sum_{j=1}^{t} \left[b^2_{(jk)j.k}\ a_j^2 - b^2_{(jk)j.k}\ (1 - a_j^2)\ \frac{a_j^2}{1 - a_j^2} \right] .$$

That is for each of the m sets of trait weights the two factor results squared should be related to the h^2 results by a slope parameter, $\frac{a_j^2}{1 - a_j^2}$.

Similarly for each of the t sets of method weights the two factor results squared should be related to the $\hat{h}^2$ results by a slope parameter, $\dfrac{a_k^2}{1 - a_k^2}$. The two factor results can be rotated so that this linear condition with zero intercept is most closely realized. This could be done in several ways, either differences or ratios in combination with either least square or maximum likelihood. However, if the model fits and the sample is large, little difference between these prodedures will occur. However $\dfrac{a^2}{1 - a^2}$ is estimated, these estimates lead to estimates of the a's and, in turn, to estimates of the b's, and finally to the answers to the questions asked originally.

Even if this model does not fit a large amount of data, it does provide a framework by which the researcher can inspect matrices such as these. For example, Campbell and O'Connell (1965) observe that when the hetero-hetero values are plotted against the homomethod-heterotrait values, the relationship appears linear but the latter set of values are more variable. That is,

$$\sigma^2_{r_{(jk)(j'k)}} > \sigma^2_{r_{(jk)(j'k')}}$$

Using the simplest model or the model which assumes trait variance independent of method variance, it becomes obvious why this empirical result occurs. That is, for the simplest model

$$r_{(jk)(j'k)} = b_{(jk)j.k}\, b_{(j'k)j'.k}\, r_{jj'} + b_{(jk)k.j}\, b_{(j'k)k.j'}\, (1.00)$$

and

$$r_{(jk)(j'k')} = b_{(jk)j.k}\, b_{(j'k')j'.k'}\, r_{jj'} .$$

It is obvious from inspecting these two equations that the first equation has two sources of variance whereas the second equation has only one source of variance. One would expect the second equation to be more variable than the first equation only if one (or both) of the methods had little variance (i.e., $b_{(jk)k.j} \rightarrow 0$) and the trait weights for variables using that method were more variable than the trait weights for the other method (i.e., $b_{(j'k)j'.k} > b_{(j'k')j'.k}$).

References

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, D. T., & O'Connell, E. J. Methods factors in multitrait-multimethod matrices: Multiplicative rather than additive? Duplicated contract report, Northwestern University, 1965.

Methods Factors in Multitrait-Multimethod Matrices: Multiplicative rather than Additive?[1]

Donald T. Campbell and Edward J. O'Connell
Northwestern University Syracuse University

---

[1] The preparation of this paper was supported by Project C998, Contract 3-20-001, with the Educational Media Branch, Office of Education, U.S. Department of Health, Education and Welfare, under provisions of Title VII of the National Defense Education Act. The authors are greatly indebted to Mrs. Carole Siegman for assistance at many stages of the work.

A multitrait-multimethod matrix contains all the correlations among a set of measures in which each of several traits is measured by each of several methods (Campbell,& Fiske, 1959). Such a matrix has an a priori factorial structure. Consider the various correlations describing the interrelationships among the traits: One type of these can be designated the heterotrait-heteromethod values. In these, the variables correlated differ in both trait and method. Another type can be designated the heterotrait-monomethod values. In these the traits differ, but all are measured by a common method. The purpose of this paper is to compare the level of correlation between traits under these two conditions.

Typically in such matrices (e.g., Campbell & Fiske, 1959) the monomethod correlations run higher than the heteromethod correlations, indicating that their shared method is a source of covariance. In the typical factor analytic model, if method be conceived as a factor, then we would expect the monomethod correlations to differ from the corresponding heteromethod correlations by a relatively constant average amount. While we would not expect all variables to have the same loading on this factor, neither would we expect these method loadings to bear any systematic relation to the loadings on the trait components. Thus we have no reason to expect that the difference in magnitude between the two types of correlation would be greater for two traits that are highly related than for two traits that are more nearly independent. Except as limited by the ceiling as one approaches correlations of unity, one would expect a constant or better, an unsystematic elevation of the heterotrait monomethod values over the heterotrait heteromethod values.

To test this prediction, one needs matrices meeting these requirements:

1) There should be evidence of a methods factor, indicated by a generally higher level of correlation in the monomethod blocks. 2) There should be a qualitative confirmation of the general model in that the trait relationships should show a similar pattern in both types: i.e., if traits A and B are more highly correlated than traits C and D in monomethod correlations, this should also hold true in the heteromethod correlations also. This pattern similarity can be expressed as a correlation coefficient describing the similarity between parallel intertrait correlations between the two types of submatrices (Campbell & Fiske, 1959, pp. 99-100). 3) There should be a substantial number of traits to provide sufficient comparisons to average out sampling fluctuations. 4) The correlations should be computed on a large number of persons, to provide sample stability sufficient to make the hypothesis testable. This paper will utilize a number of studies meeting these requirements.

First, let us make more explicit the alternative hypotheses involved. Table 1 presents a hypothetical multitrait-multimethod matrix showing the correlations among traits A, B, C, and D, measured by methods 1 and 2. Method 1 contributes some covariance, elevating the heterotrait values about .10 correlation points above what they are when no shared method is involved, i.e., in the heterotrait-heteromethod triangles. Method 2 contributes even more correlation, an average elevation of .20 points. As depicted in the major portion of Table 1, the methods contributions are simply additive, the method component elevating the low .10 correlation between C_1 and D_2 as much as the high .40 correlation between A_1 and B_2.

To the right of the main table, are two alternative models, illustrated only for the A_1 column. In the first, the methods component is inversely

Table 1. Sample Multitrait-Multimethod Matrix: Methods 1 & 2, Traits A, B, C, & D

	A_1	B_1	C_1	D_1	A_2	B_2	C_2	D_2	(Alternative A_1 columns) (Non-additive method components)	
A_1	(.60)	.50	.40	.30	.50	.40	.30	.20	(.55)	(.75)
B_1	.50	(.60)	.35	.45	.41	.50	.25	.35	.50	.60
C_1	.40	.35	(.60)	.20	.29	.26	.50	.10	.45	.45
D_1	.30	.45	.20	(.60)	.21	.34	.11	.50	.40	.30
A_2	.50	.41	.29	.21	(.70)	.60	.50	.40	.50	.50
B_2	.40	.50	.26	.34	.60	(.70)	.45	.55	.40	.40
C_2	.30	.25	.50	.11	.50	.45	(.70)	.30	.30	.30
D_2	.20	.35	.10	.50	.40	.55	.30	(.70)	.20	.20

proportional to the basic trait correlation--the higher the latter, the less the correlation is augmented when both traits are measured by the same method. The column to the far right in Table 1 shows a multiplicative effect: the higher the basic, heterotrait-heteromethod correlation, the more it is elevated by shared method.

Figure 1 shows a mode of plotting the alternative effects which will be used in the subsequent empirical studies. In part (a), the simple additive relationship is plotted, using the below diagonal triangle of monomethod 1, heterotrait values only, plotted against the 6 corresponding below-validity-diagonal values of the heteromethod-heterotrait-triangle. (The validity diagonal and reliability values have been omitted because in so many studies good reliabilities are not available.) The additivity shows up as a line of values paralleling the identity diagonal, lying above it since the monomethod values are higher. In part (b) of Figure 1, the increment is inversely proportional, in part (c), multiplicative.

Before we look at empirical data, it may be well to examine which of these outcomes has the best theoretical expectation. Outcome (a), the additive, has simply the argument of independence. If methods factor loading and trait factor loadings are in fact independent, then a fuzzy version of this parallel model is our best guess. Outcome (b) is favored by a number of considerations: at the extreme, trait and method factors--as any two factors--do interact. Increasing the variance contributed by one decreases the factor loading of the others--since factor loadings express the contribution as a proportion of total variance. Looking at (a) and (c), we see that the trends cannot be extended fully or monomethod correlations larger than unity would be encountered. Eventually the plot has to slope toward the diagonal as unity is approached. If one suspects that the additivity is more likely at the level of covariances than for r, then the slope of type (b) is expected in terms of r. If one expects proportionality or equivalence of r units to be better achieved with r^2 or with z' transformations, then a slope of type (b) is to be expected when expressed in r.

To analyze empirical multitrait-multimethod matrices, a computer program (O'Connell, 1962-65) was developed to display the outcome as a scatter diagram and to compute various relevant indices. The program handles two methods, and a large number of traits. To avoid confusing a dog-lagged version of outcome (a) [shown in (d) of Figure 1] with outcome (c), the program reflects variables, based upon a pooled criterion of both monomethod triangles. This represents an effort to assure that the method factor is weighted in the same direction for all variables, moving the matrix as much as possible into the all-positive quadrant.

The computer program compares each of the two monomethod heterotrait triangles with each of the two heterotrait-heteromethod triangles (treating the values under the validity diagonal in half the comparisons, those above the validity diagonal in the others). For each of these four, a plot such as shown in Figure 2 has been prepared, computing the values shown there. The r_r stands for the coefficient of correlation computed for the relationship shown in the plot--that is, the correlation between the corresponding values in the two triangles of heterotrait coefficients. (While all values have been used in computing r_r, not all are shown in the scatter diagrams, as multiple values at the same plotting locus are represented only once.)

The magnitude of the methods component is represented in two manners, one by the "elevation," the difference in mean level between the monomethod values and the heteromethod values. The other is by the slope, which in the additive model would be 1.00. This slope is the average of the two regression lines, or the ratio of the two standard deviations.

The results from a variety of explorations are sampled in Figures 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11, and in Tables 2 and 3. These studies point overwhelmingly to a multiplicative function, to a slope of type (c) with the line of fit going through the origin, the zero-zero point, the methods contribution increasing as the magnitude of the inter-trait relations increases. While eventually such a function would have to bend to the right and become asymptotic to 1.00, within the range sampled, no evidence of this occurs.

Table 2 presents data from a number of validity studies. In column K, the number of variables are presented, N is the number of persons on whom the correlations are computed, under "Mean $\underline{r}$" are the average values of correlations in various heterotrait triangles. M_1 and M_2 stand for the two monomethod triangles. Het A represents the heteromethod heterotrait triangle lying above the validity diagonal, Het B that lying below. Under "Elevation" are presented the differences between each of the two heterotrait monomethod means and each of the two heterotrait heteromethod triangle means. Under "Sigma $\underline{r}$" are the standard deviations of coefficients in the four triangles Under "Slope" are the ratios of nonomethod over heteromethod sigmas. Under "$\underline{r}_r$" are the pattern similarity coefficients correlating corresponding heterotrait values in the two triangles. Under "Background" are slope and $\underline{r}_r$ values between the two monomethod triangles, and between the two heteromethod triangles, data not directly relevant to the question of method contribution, but reflecting on quality.

The studies are designated as follows: "Kelly-Fiske" refers to the

Table 2. Slopes and Elevations for Validity Matrices

			Mean r				Elevation				Sigma r			
	K	N	M_1	Het A	Het B	M_2	M_1-Het A	M_1-Het B	M_2-Het A	M_2-Het B	M_1	Het A	Het B	M_2
Kelly-Fiske														
Staff-Teammates	22	124	.223	.135	.132	.234	.088	.091	.099	.102	.245	.180	.194	.214
Staff-Self	22	124	.223	.050	.064	.192	.173	.159	.142	.128	.245	.130	.140	.138
Teammates-Self	22	124	.234	.069	.056	.192	.165	.178	.123	.136	.214	.123	.119	.138
Carroll														
Guilford-Martin-Peer	5	110	.442	.158	.130	.337	.284	.312	.179	.207	.255	.148	.131	.185
Guilford-Martin-Self	5	110	.442	.274	.263	.200	.168	.179	-.074	-.063	.255	.162	.141	.131
Self-Peer	5	110	.200	.205	.121	.337	-.05	.079	.132	.216	.131	.125	.082	.185
Tutko														
Self-Peer	16	100	.062	-.002	.053	.105	.064	.009	.107	.052	.156	.172	.170	.352
Self-Edwards	16	100	.053	.037	.067	.074	.016	-.014	.037	.007	.159	.126	.112	.155
Peer-Edwards	16	100	.201	.041	.049	.055	.160	.152	.014	.006	.308	.117	.113	.163
Meeland-Becker														
Peer-Questionnaire	4	102	.591	.359	.281	.268	.232	.310	-.091	-.013	.172	.114	.150	.152
Wetzel-Peterson														
Self-parent	24	72	.129	.074	.070	.156	.055	.059	.082	.086	.230	.138	.147	.240
Self-Peer	24	72	.117	.011	.019	.147	.106	.098	.136	.128	.236	.169	.173	.257
Parent-Peer	24	72	.165	.025	.023	.137	.140	.142	.112	.114	.233	.136	.146	.263
Banta-Rip Van Winkle														
Fathers-Mothers	21	500	.052	.037	.037	.073	.015	.015	.036	.036	.123	.110	.078	.122
Projection Study														
Self-Peer, Female	27	204	.225	.010	.024	.282	.215	.201	.272	.258	.218	.086	.104	.236
Self-Peer, Male	27	200	.188	.033	.055	.342	.155	.133	.309	.287	.177	.119	.131	.315
zz Self-zz Peer, F.	27	204	.067	.032	.044	.146	.035	.023	.114	.102	.134	.103	.122	.234
zz Self-zz Peer, M.	27	200	.060	.048	.058	.170	.012	.002	.122	.112	.135	.114	.114	.222
Peer Attribution-Photo Attr. F.	27	204	.462	.283	.303	.505	.179	.159	.222	.202	.242	.137	.120	.201

Table 2. Slopes and Elevations for Validity Matrices (Continued)

			Mean *r*				Elevation				Sigma *r*			
	K	N	M_1	Het A	Het B	M_2	M_1-Het A	M_1-Het B	M_2-Het A	M_2-Het B	M_1	Het A	Het B	M_2
Peer Attribution-Photo Attr. M.	27	200	.435	.317	.288	.451	.118	.147	.134	.163	.286	.124	.139	.178
Holtzman-Inkblot Individual-Group	18	100	.127	.107	.098	.165	.020	.029	.058	.067	.233	.157	.165	.225

Table 2. Slopes and Elevations for Validity Matrices (Continued)

	Slope				r_r				Background			
	M_1/ Het A	M_1/ Het B	M_2/ Het A	M_2/ Het B	M_1 Het A	M_1 Het B	M_2 Het A	M_2 Het B	M_1M_2 Slope	M_1M_2 r_r	Het A Het B Slope	Het A Het B r_r
Kelly-Fiske												
Staff-Teammates	1.36	1.27	1.19	1.10	.80	.69	.84	.79	1.15	.76	.93	.67
Staff-Self	1.89	1.75	1.06	.99	.48	.60	.49	.47	1.78	.61	.93	.42
Teammates-Self	1.73	1.79	1.12	1.16	.58	.67	.46	.45	1.55	.64	1.03	.56
Carroll												
Guilford-Martin-Peer	1.73	1.95	1.26	1.42	.44	.69	.82	.78	1.37	.62	1.13	.38
Guilford-Martin-Self	1.57	1.81	.81	.93	.62	.70	.26	.82	1.94	.71	1.15	.09
Self-Peer	1.05	1.60	1.48	2.25	.07	.21	.45	.83	.71	.29	1.52	.20
Tutko												
Self-Peer	.90	.92	2.04	2.07	.66	.69	.72	.80	.44	.72	1.01	.54
Self-Edwards	1.26	1.41	1.23	1.38	.38	.51	.56	.55	1.03	.18	1.12	.34
Peer-Edwards	2.62	2.73	1.39	1.44	.48	.51	.37	.54	1.89	.19	1.04	.41
Meeland-Becker												
Peer-Questionnaire	1.51	1.15	1.33	1.01	.65	.58	.57	.43	1.14	-.02	.76	.55
Wetzel-Peterson												
Self-Parent	1.66	1.56	1.74	1.63	.48	.60	.46	.40	.96	.52	.94	.32
Self-Peer	1.39	1.36	1.52	1.49	.56	.56	.59	.60	.92	.66	.98	.48
Parent-Peer	1.72	1.60	1.93	1.80	.39	.44	.52	.57	.89	.70	.93	.44
Bents-Rip Van Winkle												
Fathers-Mothers	1.23	1.27	1.22	1.55	.62	.73	.57	.78	1.01	.83	1.27	.42
Projection Study												
Self-Peer, Female	2.53	2.08	2.74	2.26	.58	.42	.53	.56	.92	.79	.82	.44
Self-Peer, Male	1.48	1.35	2.64	2.40	.62	.57	.66	.70	.56	.83	.91	.56
zz Self-zz Peer, F.	1.31	1.01	2.28	1.91	.50	.42	.51	.67	.57	.53	.84	.35
zz Self-zz Peer, M.	1.18	1.18	1.94	1.94	.60	.57	.70	.74	.61	.63	1.00	.55
Peer Attribution-Photo Attr. F.	1.76	2.02	1.46	1.68	.74	.83	.85	.84	1.20	.89	1.15	.68

Table 2. Slope and Elevations for Validity Matrices (Continued)

	Slope				$\underline{r}_r$				Background			
	M_1/ Het A	M_1/ Het B	M_2/ Het A	M_2/ Het B	M_1 Het A	M_1 Het B	M_2 Het A	M_2 Het B	M_1M_2 Slope	M_1M_2 $\underline{r}_r$	Het A Het B Slope	Het A Het B $\underline{r}_r$
Peer Attribution-Photo Attr. M	2.31	2.07	1.44	1.28	.71	.86	.77	.69	1.61	.82	.89	.52
Holtzman-Inkblot Individual-Group	1.48	1.41	1.43	1.36	.72	.78	.70	.82	1.04	.80	.95	.67

Table 3. Test-Retest and Reliability Matrices

			Mean r				Elevation				Sigma r			
	K	N	M_1	Het A	Het B	M_2	M_1-Het A	M_1-Het B	M_2-Het A	M_2-Het B	M_1	Het A	Het B	M_2
Vassar	9	756	.176	.127	.111	.190	.049	.065	.063	.079	.351	.236	.269	.338
Gunderson-Antarctic	10	110	.325	.269	.266	.344	.056	.059	.075	.078	.224	.168	.195	.215
Owen-Bechtoldt Army Alpha	8		.359	.327	.389	.446	.032	-.030	.119	.057	.108	.117	.089	.126
Bechtoldt-Dysinger	10	161	.346	.330	.331	.378	.016	.015	.048	.047	.195	.185	.162	.175
Projection Study														
Self, F. Test-retest	27	158	.236	.181	.176	.218	.055	.060	.037	.042	.176	.150	.146	.166
Self, M. Test-retest	27	157	.184	.133	.125	.169	.051	.059	.036	.044	.168	.140	.142	.162
Rep. Female	27	237	.312	.252	.240	.348	.060	.072	.096	.108	.257	.221	.224	.261
Rep. Male	27	241	.301	.216	.200	.314	.085	.101	.098	.114	.267	.216	.221	.255

Table 3. Test-Retest and Reliability Matrices (Continued)

	Slope				$\underline{r}_r$				Background			
	M_1/ Het A	M_1/ Het B	M_2/ Het A	M_2/ Het B	M_1 Het A	M_1 Het B	M_2 Het A	M_2 Het B	M_1M_2 Slope	M_1M_2 $\underline{r}_r$	Het A Het B Slope	Het A Het B $\underline{r}_r$
Vassar	1.49	1.31	1.43	1.26	.99	.98	.99	.98	1.04	.99	.88	.95
Gunderson-Antarctic	1.33	1.15	1.28	1.10	.94	.91	.88	.91	1.04	.89	.87	.89
Owen-Bechtoldt Army Alpha*	.93	1.21	1.08	1.41	.66	.73	.68	.57	.86	.60	1.31	.49
Bechtoldt-Dysinger	1.05	1.21	.95	1.08	.95	.96	.98	.95	1.11	.95	1.14	.91
Projection Study												
Self, F. Test-retest	1.17	1.21	1.11	1.14	.88	.89	.88	.88	1.06	.86	1.03	.81
Self, M. Test-retest	1.20	1.18	1.16	1.14	.83	.85	.82	.83	1.03	.78	.99	.75
Rep. Female	1.16	1.15	1.18	1.16	.95	.96	.94	.94	.98	.92	.99	.93
Rep. Male	1.24	1.20	1.18	1.15	.92	.94	.93	.92	1.04	.93	.97	.91

data from Kelly & Fiske (1951), as reanalyzed (Campbell & Fiske, 1959). Ratings of graduate students in clinical psychology by an assessment staff, by their assessment period peers, and self-ratings are involved. (See Table 2 and Figures 2 and 3.) The Carroll (1952) study (Table 2, Figure 4) employs the Guilford-Martin personality inventory, plus dormitory peer ratings and self ratings. The Tutko PPS study represents unpublished research by Dr. Thomas Tutko, employing the Edwards Personality Preference Schedule with the Taylor Manifest Anxiety Scale as a sixteenth measure, correlated with dormitory peer ratings on the sixteen variables, and with self ratings. The Meeland matrix is that cited by Becker (1960) involving four traits measured by peer "Behavior Ratings" and a self-descriptive questionnaire. The Wetzel study (Peterson, 1965) employs ratings by peers, self, and parents. The "Banta-Rip Van Winkle" study is from data provided by Thomas Banta from the Rip Van Winkle Child Training study conducted at Hudson, New York. Child behavior ratings are involved, with Father's ratings method 1, Mother's ratings method 2. A preliminary report on an N of 50 parent pairs has been published (Eron, Banta, Walder, & Lovlicht, 1961). The N for this matrix is variable, but around 500 on the average. From the projection project (Campbell, Miller,Lubetsky and O'Connell, 1964) come a number of matrices involving self-ratings and ratings by dormitory peers (Figure 5, Table 2). These are analyzed as simple ratings and as transformed through a "double standardization," in which all traits within raters are given the same mean and standard deviation, and these are then all standardized across persons (See Campbell et al., 1964 for details). This serves to reduce the methods component as judged by both elevation and slope. A "validity" study of attribution is also provided, relating mean attributions of traits to photographed persons and mean attributions to actual dormitory peers. The Holtzman Inkblot Study (Holtzman, Moseley, Reinehr, and Abbott, 1963) compares group and individual inkblot tests (Table 2 and Figure 6).

Table 3 describes a number of test-retest data, treated as multitrait-multimethod matrices. As is commonly noted, tests administered at the same time tend to correlate higher than those administered some time apart. This higher intercorrelation is treated as a "methods" component increasing contemporaneous correlations. The most impressive of these is the Vassar study (Webster, Freedman, & Heist, 1962; Rozelle, 1965) where freshmen personality tests are correlated with senior tests administered three or four years later (Table 3 and Figure 7). Gunderson and Nelson (1965; Rozelle, 1965) have provided us with data on 10 individual and social attitude dimensions, measured siz months apart on 110 technicians and scientists working in the Antarctic. Bechtoldt(1956) has provided two matrices for intelligence tests administered on two occasions. In the Owen-Bechtoldt matrix, the Army Alpha is readministered after 30 years. In the Bechtoldt-Dysinger matrix, only 48 hours separates the two testings (Table 3 and Figures 8 and 9). From the Projection Study (Campbell _et al_., 1964) comes a self-description task repeated two weeks later (Table 3 and Figure 10). A split half reliability study of peer ratings is also included in this table (Table 3 and Figure 11).

The preponderant finding is one of slopes greater than one. While there are a few instances of slopes less than 1.00, as one each in the Bechtoldt matrices and in the self variable for the Tutko and Carroll studies, they occur where the method component is low. In general, slope and elevation show a correlation of about .60 which remains at about that level even if only high $\underline{r}_r$ matrices are considered.

Auxiliary analyses. To examine the effect of matrix stability upon the slope coefficients, the Projection Study, Reputation X Self-description, Female, data was broken down in to eighths (n = 25), quarters (n = 51), and halves (n = 102), for comparison with the full population data (n = 204.) The outcome of this is shown in Table 4. The expected effects of matrix quality are shown in the $\underline{r}_r$ figures, which go steadily up with n. While one would expect no effect on the mean level of correlation, the smaller samples show slightly higher means, probably due to the opportunistic effects of the reflection procedure, for the n = 25 matrices reflect more variables per matrix, on the average (the final segment with n = 29 reflected 6 variables, whereas in the total 204 run, none were reflected).

Of most interest is the effect on slope: in general, the smaller the n the less the slope. This is consistently true for the average values for monomethod 1, while for M_2 the half-sample values are slightly larger than the full sample values. Certainly here is clear proof that the slopes larger than one are not by products of small samples, or poor data quality in this sense.

Included in Table 4 is another control analysis. In this, slope and intercept values are computed for each variable separately, and then subsequently averaged. In programming this analysis, the triangular orientation has been abandoned, a square matrix, symmetrical across the main diagonal used instead. For each variable, comparisons are made within its column, correlations in the monomethod section being compared with corresponding heteromethod values, omitting the diagonal reliability, and sub diagonal validity values. The averages of these are shown in the final two lines of Table 4, insofar as corresponding values are computed. The two half-samples

Table 4. Experiments With Size of N and Individual Variable Slopes

			Mean r				Elevation				Sigma r			
	K	N	M_1	Het A	Het B	M_2	M_1- Het A	M_1- Het B	M_2- Het A	M_2- Het B	M_1	Het A	Het B	M_2
Projection Study														
Self-Peer F.														
Eighths	27	25	.209	.085	.059	.454	.124	.150	.369	.395	.261	.168	.195	.288
	27	25	.289	.038	.052	.347	.251	.237	.309	.295	.301	.193	.170	.299
	27	25	.287	.035	.037	.250	.252	.250	.215	.213	.234	.174	.166	.311
	27	25	.209	.079	.105	.355	.130	.104	.276	.250	.235	.192	.202	.386
	27	25	.147	.039	.016	.376	.108	.131	.337	.360	.275	.191	.187	.300
	27	25	.215	-.034	.031	.307	.249	.184	.341	.276	.243	.184	.194	.285
	27	25	.230	.034	.083	.326	.196	.147	.292	.243	.288	.214	.207	.310
	27	25	.190	.045	.040	.363	.145	.150	.318	.323	.266	.206	.228	.361
Average			.254	.040	.053	.347	.182	.169	.307	.294	.263	.190	.194	.317
Fourths	27	51	.242	.039	.036	.346	.203	.206	.307	.310	.249	.136	.135	.283
	27	51	.245	.057	.087	.263	.188	.158	.206	.176	.213	.131	.141	.314
	27	51	.182	.010	.025	.339	.172	.157	.329	.314	.206	.118	.126	.258
	27	51	.204	-.016	.002	.292	.220	.202	.308	.290	.261	.155	.161	.347
Average			.218	.022	.037	.310	.196	.181	.287	.272	.232	.135	.141	.300
Halves	27	102	.241	.048	.054	.308	.193	.187	.260	.254	.214	.103	.114	.275
	27	102	.210	-.010	.008	.304	.220	.202	.314	.296	.200	.095	.105	.284
Average			.225	.019	.031	.306	.206	.194	.287	.275	.207	.099	.109	.279
Whole	27	204	.225	.010	.024	.282	.215	.201	.272	.258	.218	.086	.104	.236
Individual Variable														
Slope Study	27	102	.250	.053	.053	.320	.197			.267	.188	.098	.090	.232
	27	102	.218	.001	.001	.315	.227			.327	.184	.099	.094	.260

Table 4. Experiments With Size of N and Individual Variable Slopes (Continued)

	Slope				$\underline{r}_r$				Background			
	M_1/ Het A	M_1/ Het B	M_2/ Het A	M_2/ Het B	M_1 Het A	M_1 Het B	M_2 Het A	M_2 Het B	M_1M_2 Slope	M_1M_2 $\underline{r}_r$	Het A Het B Slope	Het A Het B $\underline{r}_r$
Projection Study												
Self-Peer F.												
Eighths	1.55	1.34	1.71	1.47	.21	.18	.17	.25	.91	.56	.86	.09
	1.57	1.77	1.55	1.75	.46	.32	.18	.02	1.01	.57	1.13	.28
	1.35	1.41	1.79	1.88	.18	.19	-.08	.15	.75	.43	1.05	.19
	1.22	1.16	2.01	1.91	.39	.34	.33	.53	.61	.42	.95	.12
	1.44	1.47	1.57	1.61	.19	.07	.12	.09	.92	.31	1.02	-.01
	1.32	1.25	1.55	1.47	.14	.30	.05	.11	.85	.45	.95	.23
	1.35	1.39	1.45	1.50	.42	.17	.35	.50	.93	.47	1.04	.14
	1.29	1.16	1.75	1.58	-.13	-.11	-.33	-.10	.74	.30	.90	.19
AVerage	1.39	1.37	1.67	1.65	.23	.18	.10	.19	.84	.44	.99	.15
Fourths	1.84	1.85	2.08	2.10	.41	.34	.41	.26	.88	.75	1.01	.39
	1.63	1.51	2.41	2.23	.44	.46	.34	.48	.68	.71	.93	.27
	1.75	1.63	2.19	2.04	-.00	.21	.10	.15	.80	.50	.93	.25
	1.68	1.62	2.23	2.16	.42	.25	.36	.31	.75	.56	.97	.32
Average	1.72	1.65	2.23	2.13	.32	.31	.30	.30	.78	.63	.96	.31
Halves	2.08	1.88	2.68	2.42	.50	.50	.54	.56	.78	.85	.91	.54
	2.11	1.90	2.99	2.70	.37	.34	.35	.38	.70	.61	.90	.44
Average	2.09	1.89	2.83	2.56	.43	.42	.44	.47	.74	.73	.90	.49
Whole	2.53	2.08	2.74	2.26	.58	.42	.53	.56	.92	.79	.82	.44
Individual Variable												
Slope Study	2.05			2.83	.45			.46				
	2.23			3.32	.30			.31				

of 102 each have been run on this. (For these there is no Het A and Het B, rather, under Het A is listed the values computed with M_1, under Het B those computed with M_2. For means these are equal. For sigmas, not quite identical. For some reason not yet understood, the average elevations do not equal the difference in the average $\underline{r}$'s for monomethod and heteromethod segments. For elevation, slope, and $\underline{r}_r$, there is but a single value per variable, providing one average per method.) The average of these values agree quite well with those computed (3 and 4 lines above in Table 4) in the all-variables pooled procedure. If anything, the slopes are somewhat larger, ending any question that the slopes previously observed were due to certain variables providing high $\underline{r}$'s, others providing the low ones. In terms of the variable-by-variable slopes, in the first set there ranged from 1.14 to 4.06. In the second set, they ranged from .90 to 4.98, with only one value below 1.00.

Discussion

We believe that this investigation provides an empirical finding quite general in nature and quite unexpected in terms of factor analytic considerations. This finding may be limited to methods factors, or it may hold for factor analysis more generally. In the paragraphs which follow we explore certain alternative interpretations.

A "valid" component in rater's theories of personality. In our initial explorations all of our illustrations came from ratings as method, either self-ratings or ratings of others. When one of the first exceptions we found was in the Bechtoldt-Dysinger study, e.g., Figure 9, it seemed likely that the finding was specific to ratings. (Actually, the average of the Bechtoldt slopes is greater than one, with a magnitude reflecting a small methods or occasions factor.) Our theory for a rating-specific effect hypothesized a functional interaction between "true" level of trait correlation and magnitude of rater bias. Each rater is presumed to have an implicit theory of personality, e.g., Cronbach,(1958) which leads him to expectations of cooccurrance of certain traits or synonymity of certain personality descriptive terms. This implicit theory of personality, if generally shared, produces correlations among traits as rated over and above any "true" correlations between the traits. Now it is hypothesized that any generally shared implicit theory of personality will not be entirely fortuitous or arbitrary, but will rather reflect in part on reality. It is only the very striking "real" associations of traits that will be generally noticed. Being represented in the implicit theory of personality will lead to the exaggeration of that association, i.e., a high loading on the methods factor. On this basis, the interaction between methods enhancement of a correlation and its base level is functionally explained. Relationships large enough to get noticed get exaggerated. Unfortunately, this theory is quite specific to human processes involved in ratings. It does not help explain the Holtzman inkblot study nor the several test-retest findings.

Differential attenuation of high correlations. Our original orientation towards method factors was to see them as *adding* an irrelevance to something more valid in the trait factors. And certainly some of the evidence of halo effects and response sets would lead one that way. But there is an alternative, of seeing the hazard of hetero measurement as attenuating, reducing, trait-based relationships better represented when method is "held constant" rather than varied.

This point is most easily seen in the context of test-retest matrices where we have treated synchrony as contributing an irrelevant method-like covariance. Certainly there are senses in which tests administered together share correlated errors. For example, if half of the recruits being tested come in drunk and respond randomly, this will increase the range of scores on most ability and attitude tests in a way that will cause them to intercorrelate more highly than they would otherwise, an irrelevant increment in correlation. (Note that if all come in drunk, the intercorrelations will be reduced.)

But if we consider a magnificently stable pattern of interrelationships such as the Vassar data (Table 3 and Figure 7), spanning some three to four years, it seems inevitable that the slings and arrows of outrageous fortune over the passage of time will have disturbed and disrupted the delicate relationships when measured across time. Consider the effect of misidentifying two tests sets from different persons as coming from the same person. This certainly happens in many panel studies even if not in the Vassar study. Such an event would not be apt to disrupt or change a low correlation, might be as likely to raise as to lower a zero correlation, but would certainly lower a high one in the cross-time computation of it. Similarly, personality and attitude changes will lower high heterotrait heterotemporal correlations more than low ones, in comparison with the corresponding heterotrait monotemporal ones.

The heterotemporal values are attenuated, perhaps as described by the test-retest reliability values which occupy what is usually called the "validity" diagonal. Notice that the traditional correction for attenuation acts in a multiplicative way, increasing low correlations less than high ones.

Our slope values might turn out to be $1/\sqrt{\underline{r}_{11}}$ when computed on individual variables. The monotemporal matrix might just be the heterotemporal matrix corrected for attenuation. Since this would put the monotemporal reliabilities up to unity, more likely the correction is in terms of the difference between the internal consistency monotemporal reliability and the heterotemporal test-retest reliability. Since these are determinable independently of the slopes here studied, empirical examination of the convergence of the model on a trait-by-trait basis should be undertaken. An iterative improvement on the variable-by-variable slopes could be made disregarding differences in the trait composition of A_1 and A_2 (reasonable in a test-retest matrix), taking the slopes as first approximations, interpreting the slopes of B_1, C_1, D_1, etc. as known $1/\sqrt{\underline{r}_{11}}$'s, and solving for the reliability of A_1 from the difference between the $\underline{r}_{A_1B_1}$ and $\underline{r}_{A_1B_2}$. This would be done in turn for each variable, getting new estimates. Or one could begin directly with the differences in correlation pairs, solving for the reliabilities, assuming them equal and assuming $\underline{r}_{A_1B_1}$ to be the corrected-for-attenuation version of $\underline{r}_{A_1B_2}$. Averaging such first estimates for each variable, one could then assume all but one variable known to be these values, and solve for the other, etc. getting a new set of estimates, and so on iteratively.

This mode of thinking can be extended to the rating situation. Consider two raters each reporting on the same set of persons. We compare the correlation between traits A and B within one rater, A_1B_1, with the correlation across raters A_1B_2 or A_2B_1. The higher is A_1B_2 the more apt it is to be disturbed by the differences in functional definition each has for each trait, by the differences in the behavior samples of the ratees which each has observed or heard related, and by differences in interpreting meaning and intent in behavior samples both have observed. This is attenuation by partially unshared referents, and perhaps the "validity" coefficients, or agreement on the same trait, e.g., A_1A_2, is properly used as a reliability coefficient for such purposes. Similarly, the differences between group vs. individual inkblot testing could be interpreted in terms of an attenuation model.

Basic inappropriateness of the additive factor analysis model. The above two considerations make our method factor findings exceptional, make the non-additivity a special case. It must also be considered that what we have found here might be true of the factorial situation in psychology in general, not previously noted because of the lack of appropriate bodies of data.

Factor analysis occupies an atypical spot in the mathematical models of science. In most sciences the particular form of the final formula has emerged as a product of an intimate and intense interaction with the data, in which the data have in some manner been used to edit earlier formulations, to reject and select among alternatives. Often there are particular details of the formula reflecting empirical facts that made inapplicable an earlier simpler formula. None of this is true for factor analysis. The basic string of additive correlation units produced by the products of factor loadings has remained a constant, even though modes of extracting and rotating have shifted. The choice of these analysis procedures has been based upon logical and feasibility considerations, rather than on the fit of data. Basically, our additive model of factor analysis, and our choice of r rather than r^2, z', cross-products, covariances, etc. for this additive decomposition, have gone untested. We have always used factor analysis as the criterion to which data should fit, never vice versa. Seen in this light, the very nature of the factor analytic formula looks suspect. Long series of additive relationships are much rarer in the successful formulae of the successful sciences than are multiplicative and exponential relationships of great complexity. We must remember that we inherited this basic form not through any regular data series suggesting this componential model, but rather through a pragmatic decision that the actual functional relationship was probably too complex for us to hypothesize and check. In this situation, we resorted to a Taylor theorem positing that any complex function could be imitated by a sufficiently long additive series. We have inherited our model as a carryover from a day in which applied psychologists were more interested in imitating complex end products than in imitating functional interrelationships among diagnosable components. In the current usage of factor analysis this is no longer our goal--we now have an interest in delineating components and modes of interaction which is greater than our interest in "reproducing" the matrix of correlations--but the latter was the goal for which the model was chosen, and it was chosen as a part of a council of defeat as to the former goal. We have forgotten this history in our current unquestioning acceptance of the relevance of factor analysis for our analytic theoretical interests in the composition of individual differences.

The present data provide a set of instances for which factor analysis would give a systematic misfit. Applied to these data, and rotated to provide methods factors, the residuals would be systematically positive for the high intercorrelations and negative for the low intercorrelations in a way going beyond regression artifacts. The same would be true for a direct iterative solution for methods and trait factors, of a sort to be presented shortly in another report. A multiplicative model, in which the methods factor multiplied the trait factor would, we believe give a better fit.

The case for factor analysis in general can only be made by moving out of the reliability-validity domain into settings which are similar in the symmetrical crossing of two classes of factors, but which deal with contents more typically factorial. If in these settings, this same multiplicative relationship is found, the challenge to factor analysis will be genuine. It is only a conjecture until then.

Guilford's bidimensional sets of creativity tests, in which several functional modes are crossed with several problem contents, might provide the grist for such examination. Triandis' studies in which a variety of attitudes are expressed toward a variety of targets provides such. Osgood's semantic differential studies could generate such matrices, in which one would intercorrelate all of the sixty ratings made when thirty scales were applied to two concepts such as mother and father. Do the heteroparent (heteroconcept) correlations differ from their corresponding monoparent values in this multiplicative way?

Visually inspecting coefficients from studies in which five attitude dimensions (social distance, blaming, morality, capability, and affection) were each used to measure attitudes toward five outgroups (Negro, Japanese, Jewish, Mexican, and English) (Campbell, 1946; Campbell & McCandless, 1951), fails to give completely clearcut indications. For the attitude dimension components, additive or inverse relationships seem as frequent as multiplicative. For the specific outgroup components, the multiplicative relationship seem to be the rule, where there is enough range of coefficients to produce a slope.

Summary

When traits are measured by the same method (or at the same time) the intertrait correlations are higher than when the intertrait correlation is across methods. An empirical investigation of the nature of such method factors is reported, in which the method factors seem to operate in a multiplicative rather than additive way. An implicit-theory-of-personality explanation is offered for the case of ratings as methods. An attenuation model has more general applicability to the cases illustrated. The possibility is raised that the finding, if confirmed in other domains, demonstrates a fundamental inappropriateness of factor analysis for the componential analysis of individual differences data.

References

Bechtoldt, H. P. Factor analysis of test-retest data. Paper presented at the Midwestern Psychological Association Meetings, St. Louis, Mo., May 5, 1956, dittoed.

Becker, W. C. The matching of behavior rating and questionnaire personality factors. *Psychological Bulletin*, 1960, 57, 201-212.

Campbell, D. T. The generality of a social attitude. Unpublished Ph.D. dissertation, University of California, Berkeley, 1947.

Campbell, D. T., & McCandless, B. R. Ethnocentrism, xenophobia and personality. *Human Relations*, 1951, 4, 185-192.

Campbell, D. T.,& Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.

Campbell, D. T., Miller, N., Lubetsky, J., & O'Connell, E. J. Varieties of projection in trait attribution. *Psychological Monographs*, 1964, 78, No. 15, Whole No. 592, pp. 1-33.

Carroll, J. B. Ratings on traits measured by a factored personality inventory. *Journal of Abnormal and Social* Psychology, 1952, 47, 626-632.

Cronbach, L.J. Proposals leading to analytic treatment of social perception scores. In R.Tagiuri & L. Petrullo, Eds. *Person Perception and Interpersonal Behavior*. Stanford: Stanford University Press, 1958, pp 353-379

Eron, L. D., Banta, T. J., Walder, L. O., & Laulicht, J. H. Comparison of data obtained from mothers and fathers on childrearing practices, and their relation to child aggression. *Child Development*, 1961, 32, 457-472.

Gunderson, E. K. E., & Nelson, P. D. Attitude changes in small groups under prolonged isolation. Report No. 62-2, U.S.N. Med. N.P. Res. Unit, San Diego, March, 1962. Gunderson, E. K. E., & Nelson, P. D. Measurement of group effectiveness in natural isolated groups. *Journal of Social Psychology*, 1965, 66, 241-249.

Holtzman, W. H., Mosely, E. C., Reinehr, R. C., & Abbott, E. Comparison of the group method and the standard individual version of the Holtzman Inkblot Test. Journal of Clinical Psychology, 19, 433-453.

Kelly, E. L., & Fiske, D. W. The prediction of performance in clinical psychology, Ann Arbor: University of Michigan Press, 1951.

O'Connell, E. J. Fortran computer programs for summarizing multitrait-multimethod matrices. 1962-1965.

Peterson, D. R. Scope and generality of verbally defined personality factors. Psychological Review, 1965, 72, 48-59.

Rozelle, R. M. Causal relations in attitude change as demonstrated through the cross-lagged panel correlation. Mimeographed research report, Department of Psychology, Northwestern University, Evanston, Illinois, 1965.

Webster, H., Freedman, M., & Heist, P. The effects of college education. In R. N. Sanford, (Ed.), The American College. New York: Wiley, 1961, pp. 811-847.

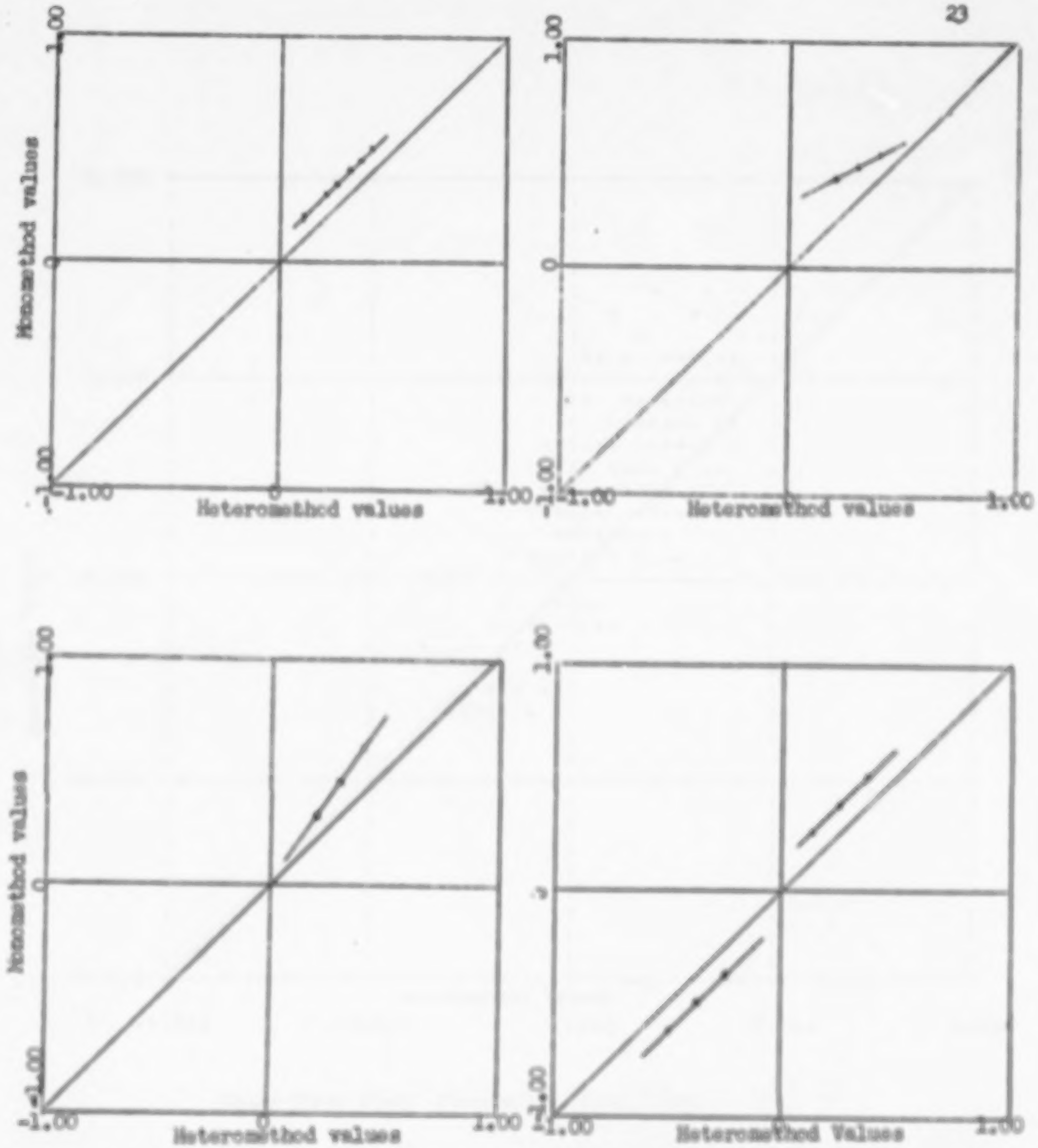
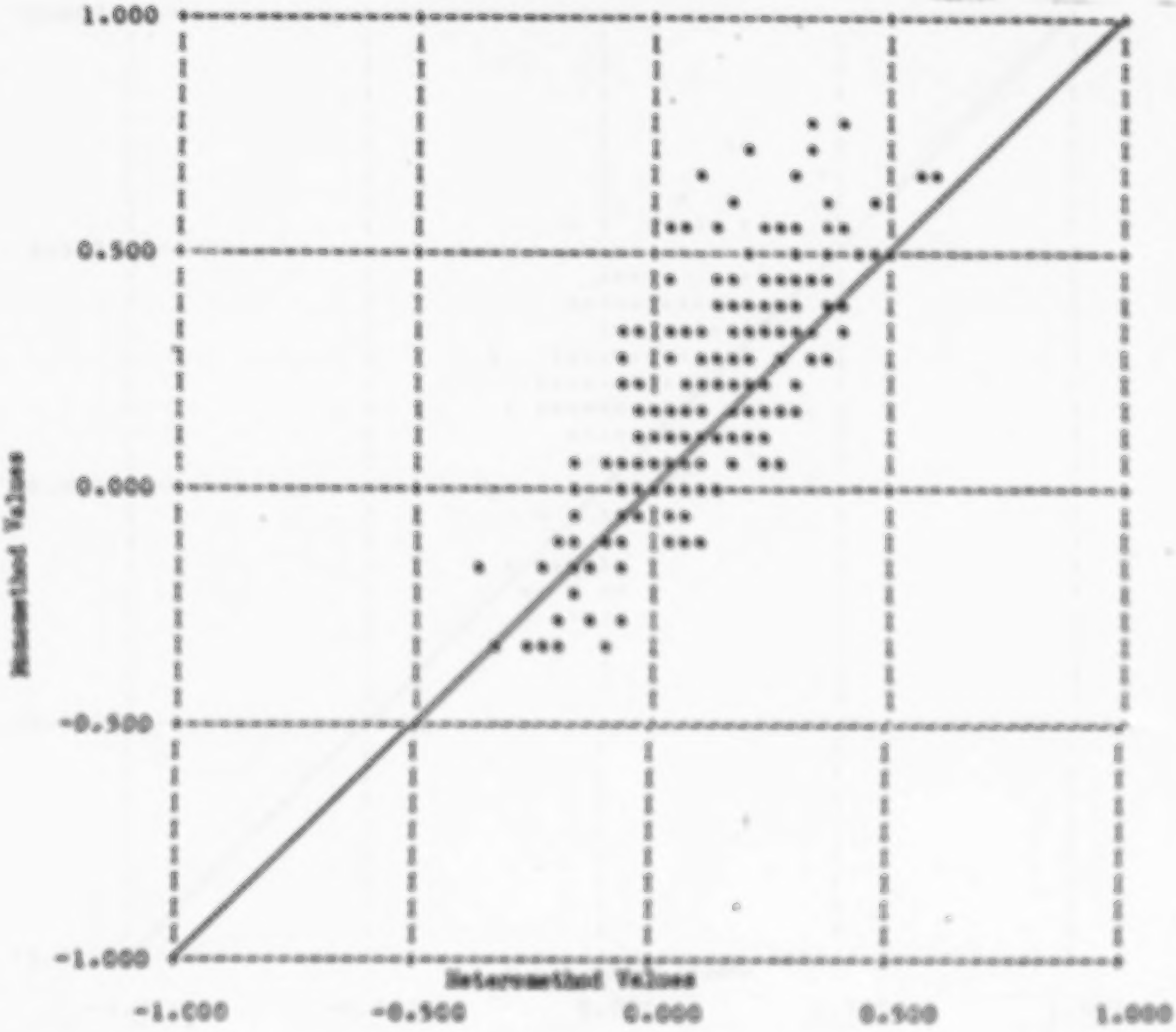


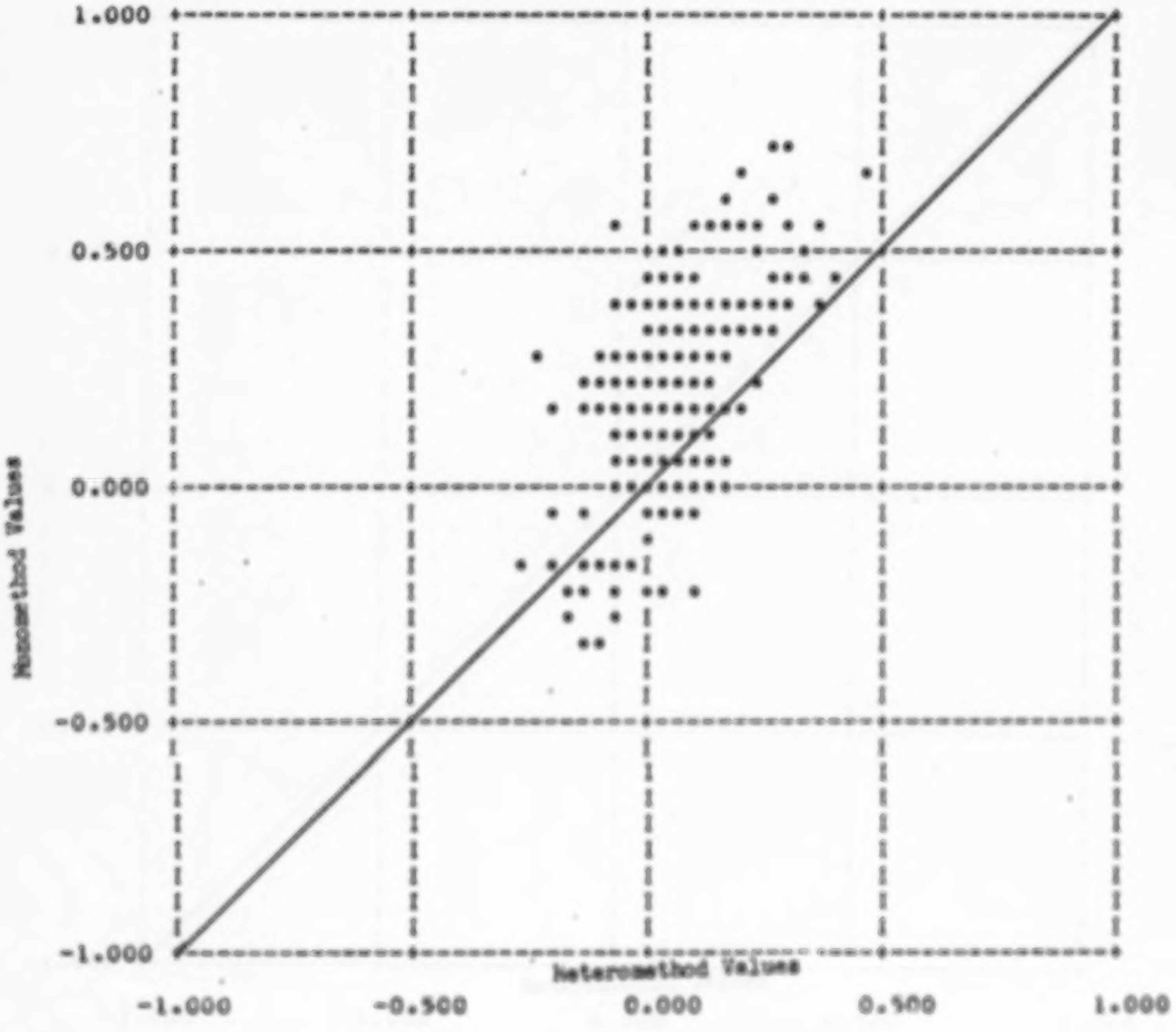
Figure 1. Hypothetical outcomes of a plot of heterotrait-*monomethod* correlations against the corresponding heterotrait-*heteromethod* values.



Kelly-Fiske study (Campbell & Fiske, 1959)

	Mean r	Elevation	Sigma r	Slope	r_p
Monomethod: Staff	.443		.245		
Heteromethod: Staff x Team	.135		.180		
		.308		1.36	.80

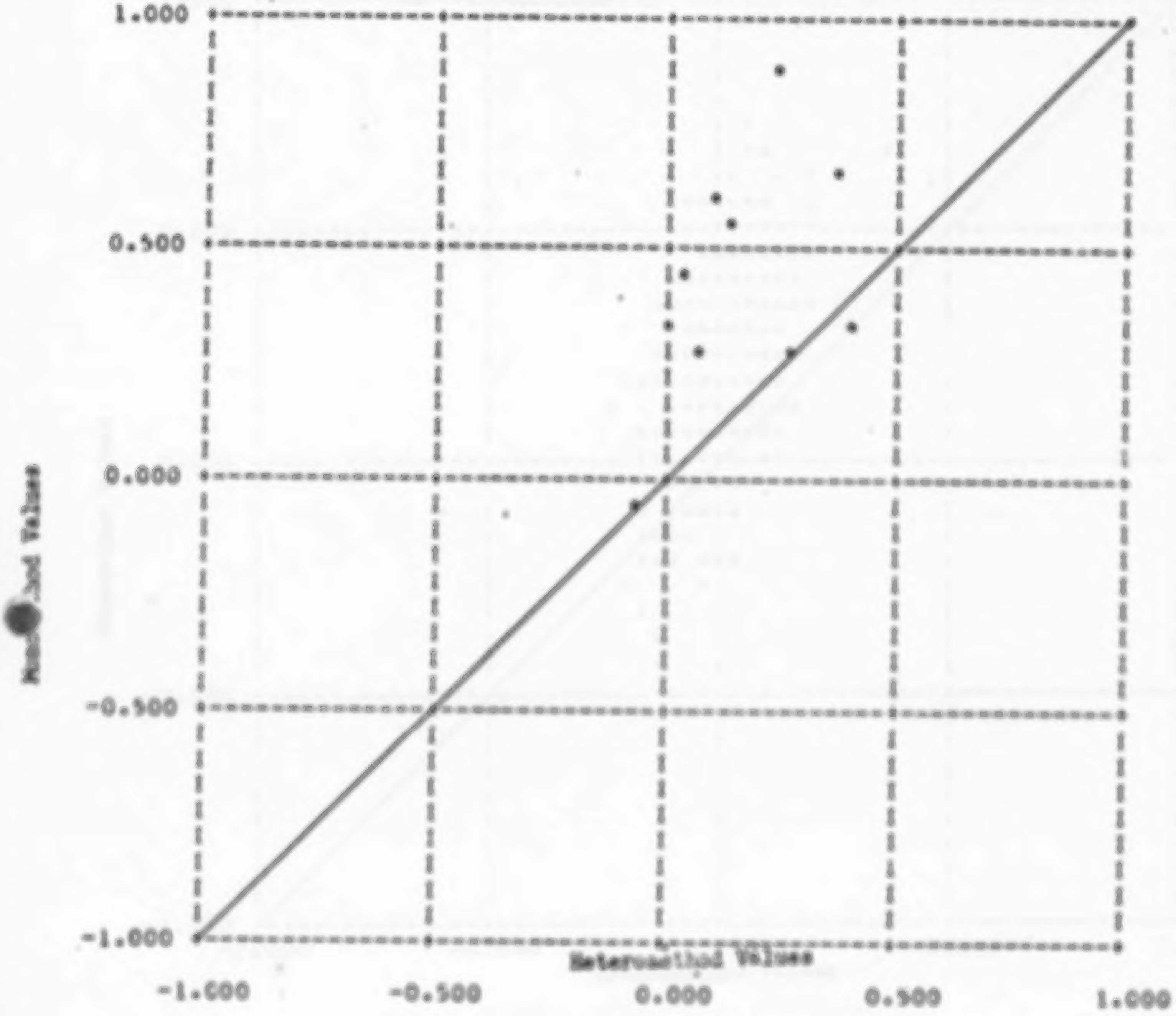
Figure 2



Kelly -Fiske study (Campbell & Fiske, 1959)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Team	.234		.214		
Heteromethod: Team x Self	.069		.123		
		.165		1.73	.58

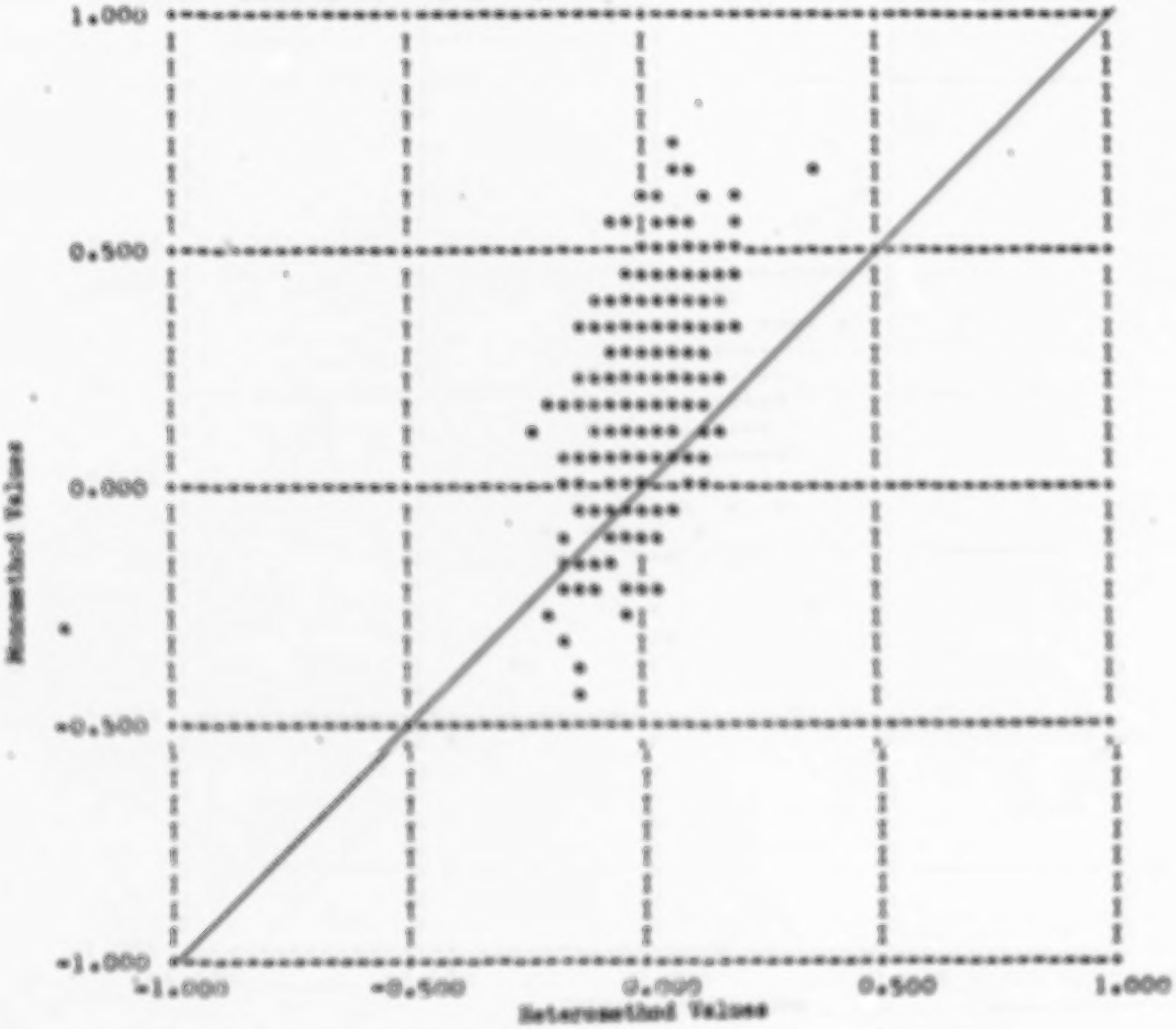
Figure 3



Guilford-Martin study (Campbell & Fiske, 1959)
(Carroll, 1952)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Inventory	.442		.255		
Heteromethod: Inventory x Peer	.158		.148		
		.284		1.73	.44

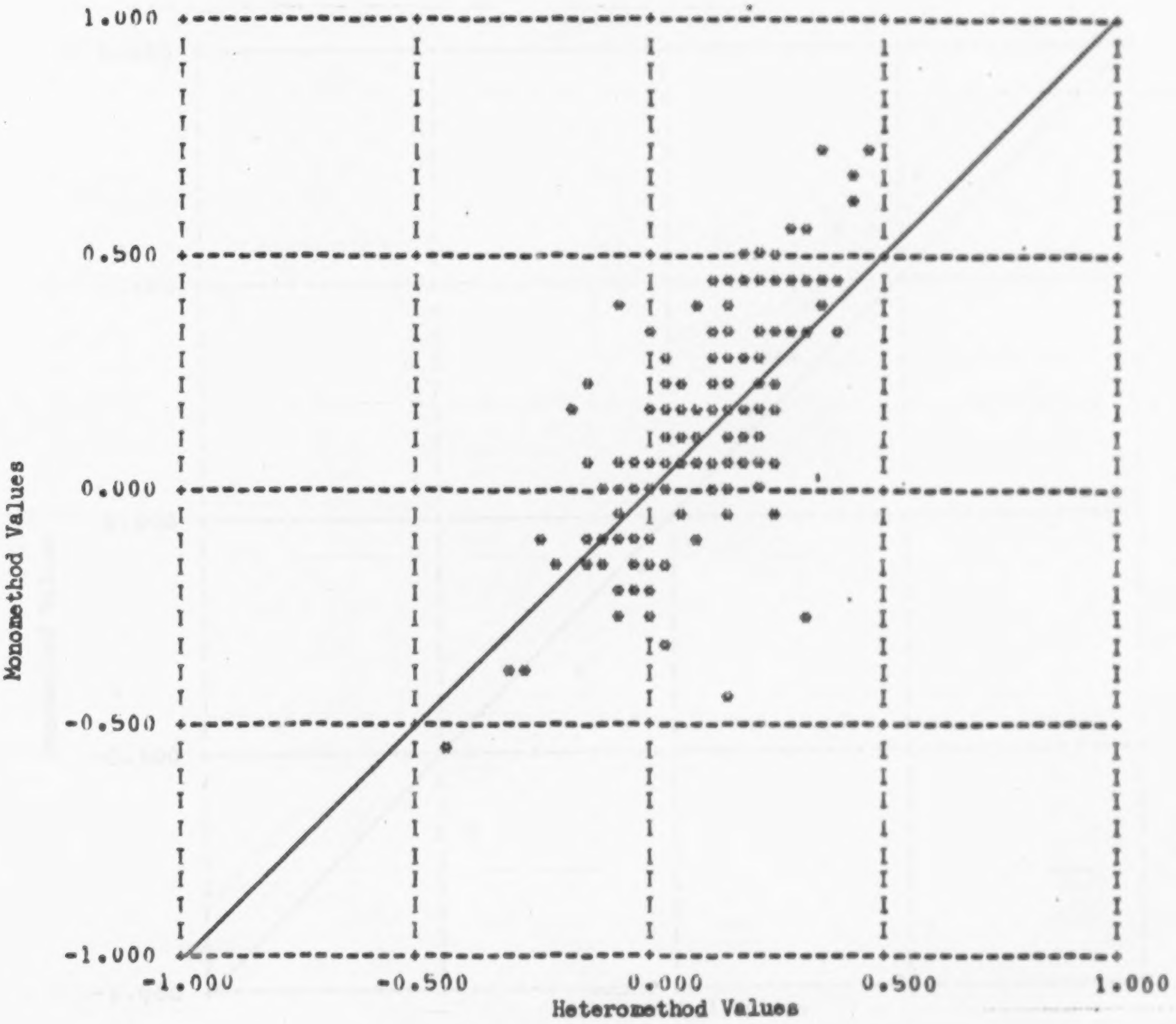
Figure 4



Projection Study (Campbell, Miller, Lubetsky, and O'Connell, 1964)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Self description raw female	.225		.218		
Heteromethod: Self description raw female x reputation raw female	.010		.086		
		.215		2.53	.58

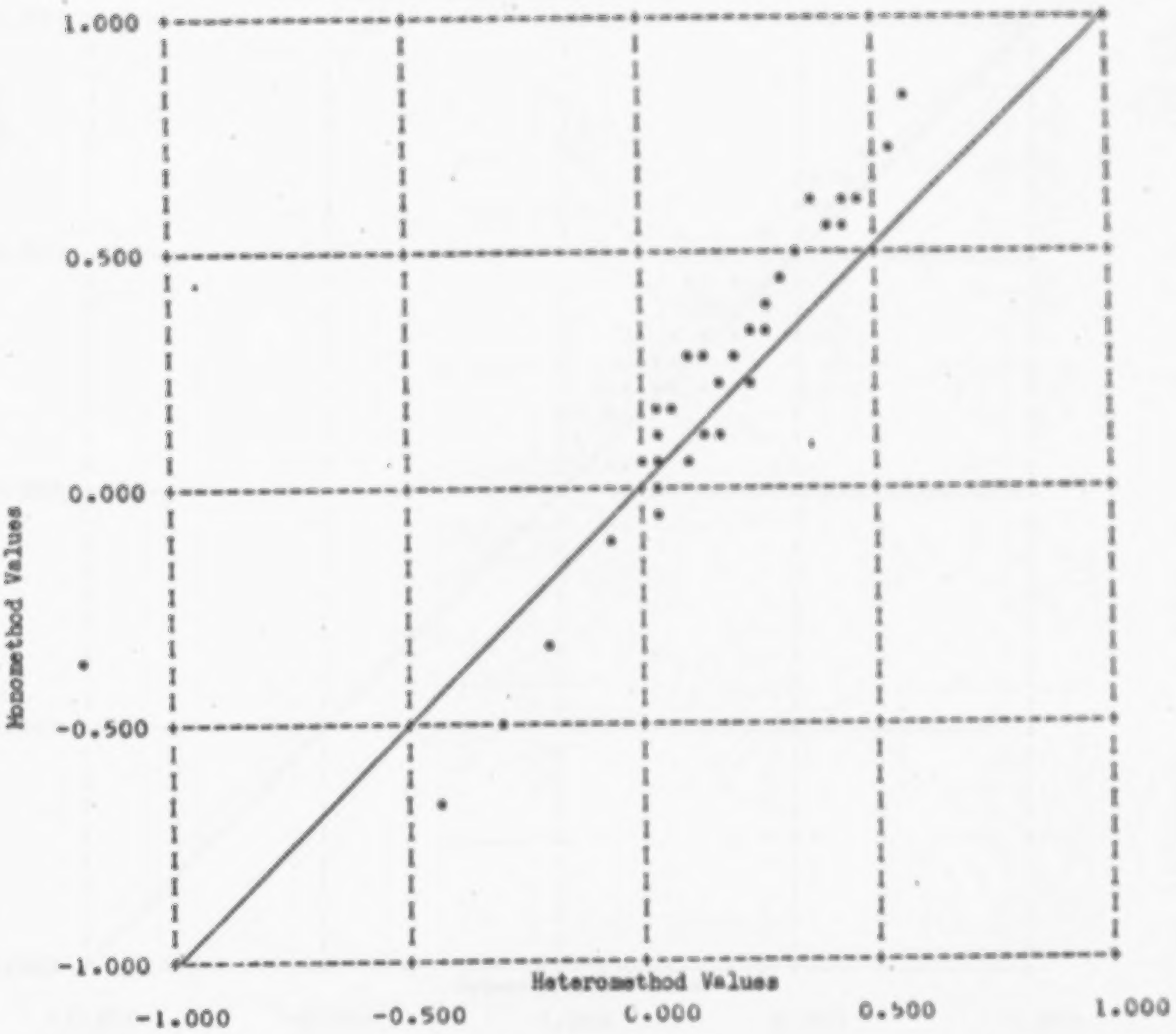
Figure 5



Holtzman study (Holtzman, Moseley, Reinehr, & Abbott, 1963)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Individual	.127		.233		
Hetercmethod: Individual x Group	.107		.157		
		.020		1.48	.72

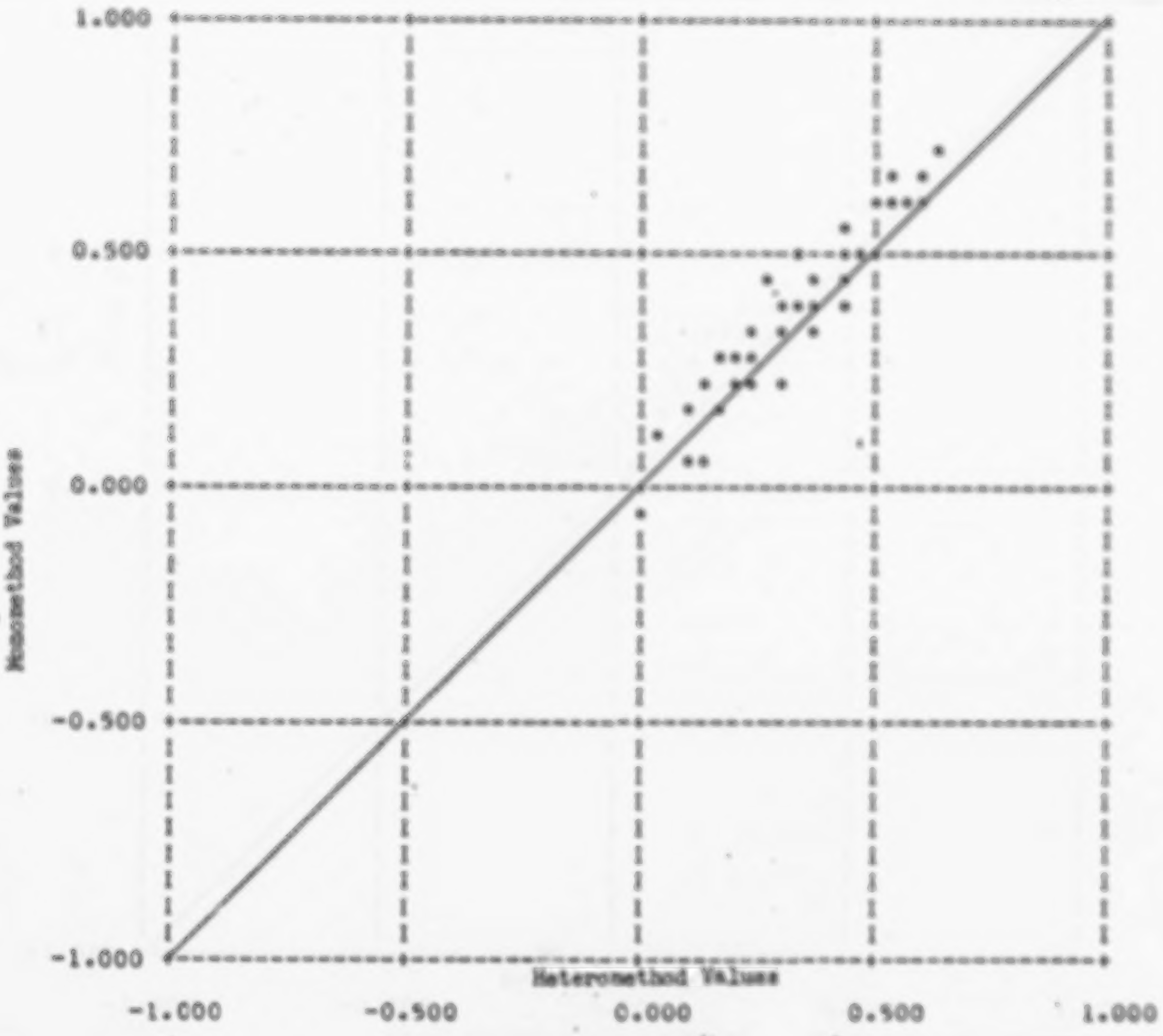
Figure 6



Vassar study (Webster, Freedman, & Heist, 1962)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Senior	.176		.351		
Heteromethod: Senior x Freshman	.127		.236		
		.049		1.49	.99

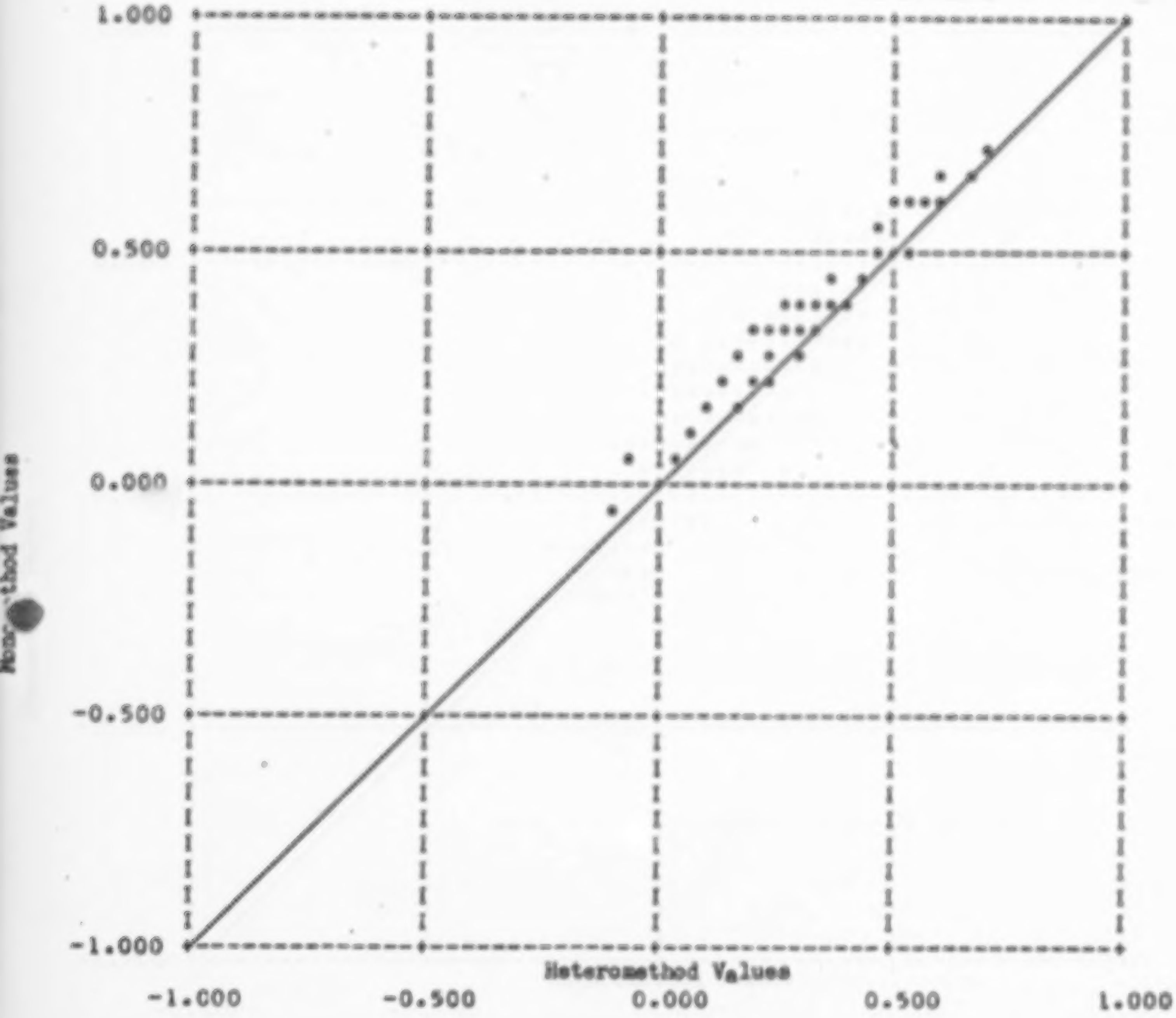
Figure 7



Bechtoldt-Dysinger study (Bechtol , 1956)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Test	.378		.175		
Heteromethod: Test x Retest	.331		.162		
		.047		1.08	.95

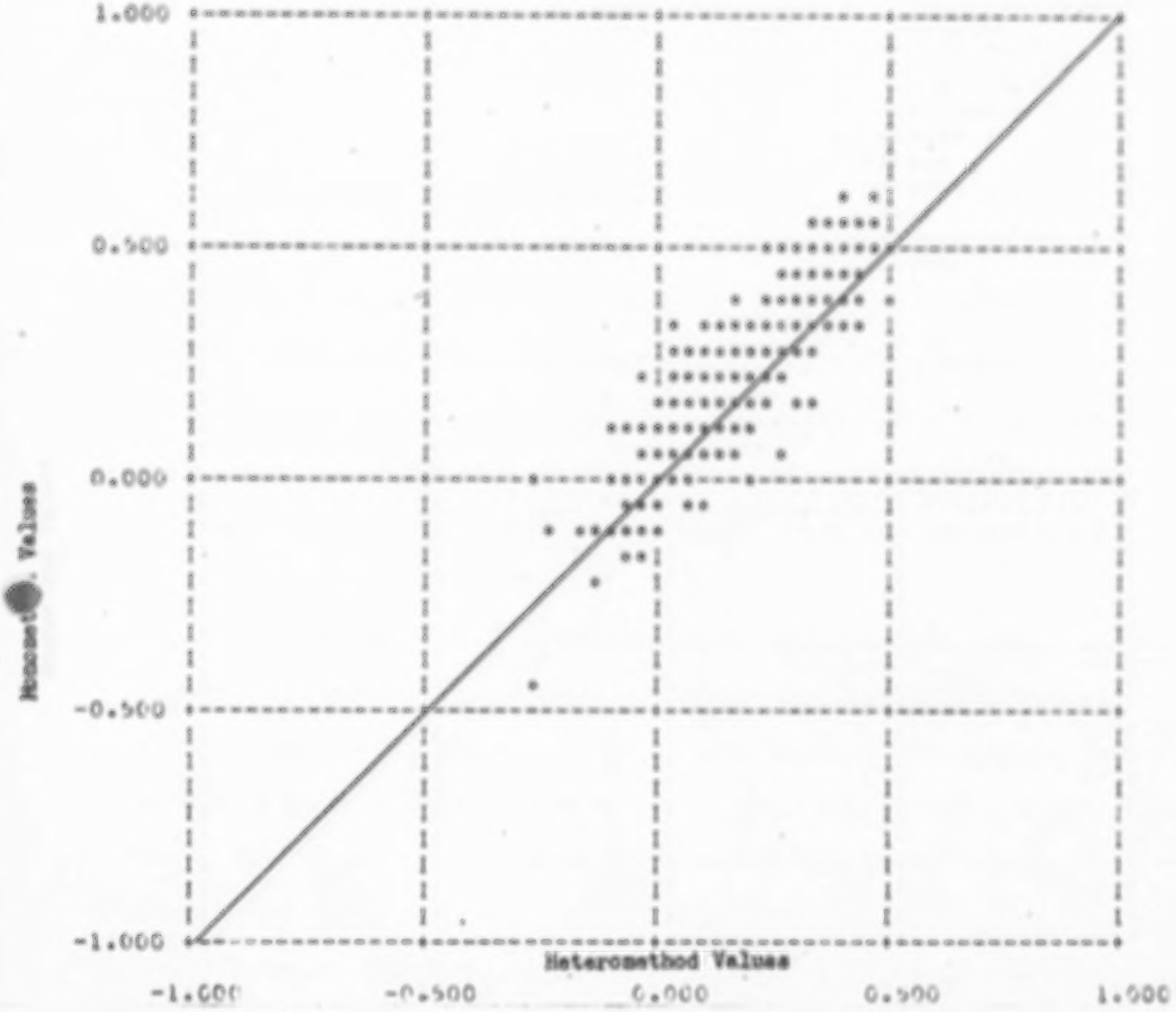
Figure 8 .



Bechtoldt-Dysinger study (Bechtoldt, 1956)

	Mean r	Elevation	Sigma r	Slope	r_r
Monomethod: Test	.378		.175		
Heteromethod: Test x Retest	.330		.185		
		.048		.95	.98

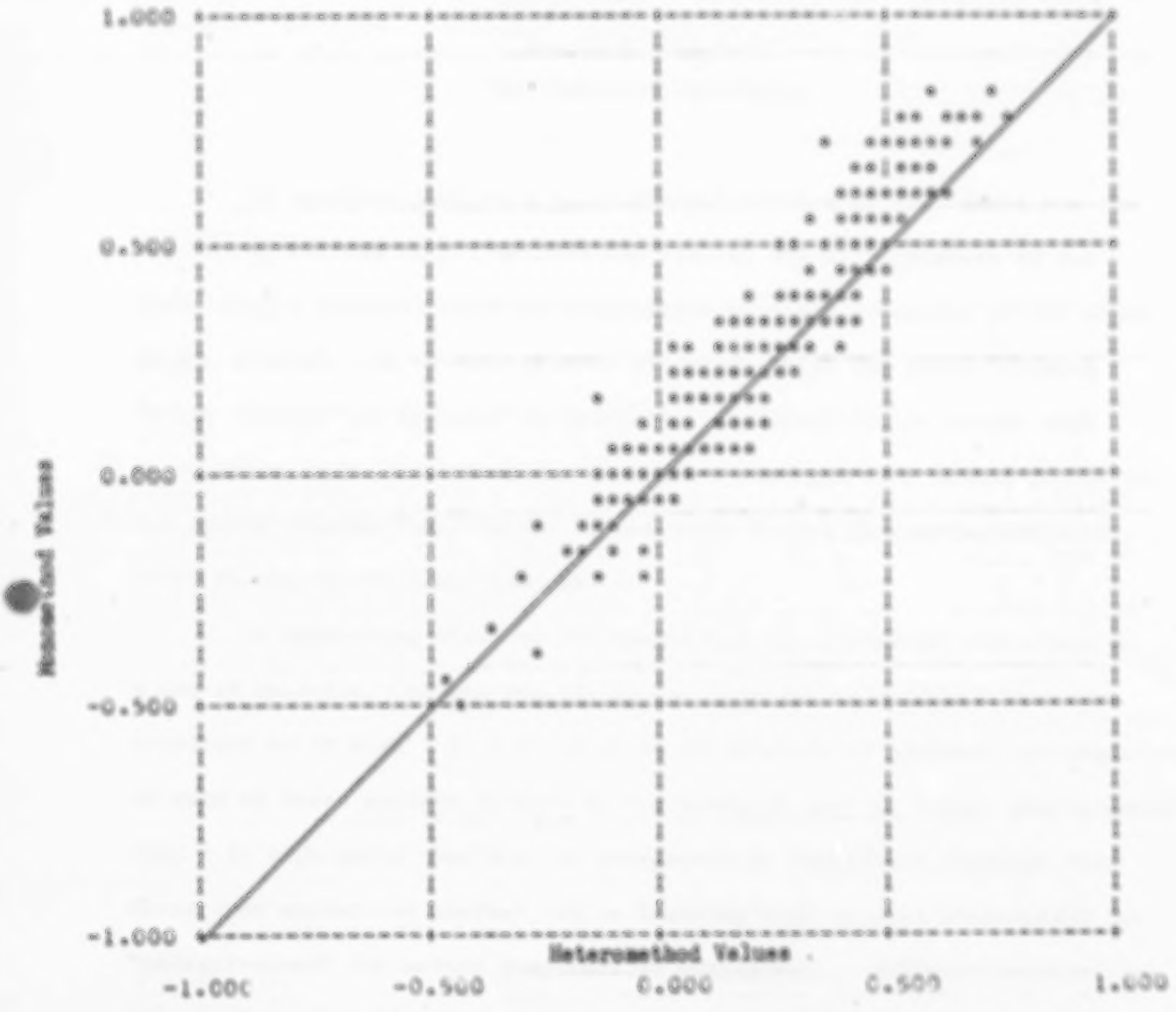
Figure 9



Projection Study (Campbell, Miller, Lubetsky, and O'Connell, 1964)

	Mean r	Elevation	Sigma r	Slope	r_r
Female raw self description					
Monomethod: Time 1	.236		.176		
Heteromethod: Time 1 x Time 2	.181		.150		
		.055		1.17	.88

Figure 10



Projection Study (Campbell, Miller, Lubetsky, and O'Connell, 1964)

	Mean r	Elevation	Sigma r	Slope	r_r
Female raw reputation					
Monomethod: Odd raters	.312		.257		
Heteromethod: Odd x even raters	.252		.221		
		.060		1.16	.95

Figure 11

DIRECT ITERATIVE FACTORING IN MATRICES WITH CROSSED FACTOR-CLASSES[1]

Donald T. Campbell
Northwestern University

In matrices such as a mulitrait-multimethod matrix, there are two classes of factors (e.g., methods and traits) and all instances of one class (e.g., methods) occur in combination with all instances of the other (e.g., traits). It is this general situation which the phase "crossed factor classes" is intended to indicate. Not every factor occurs with every other (e.g., method factor I is never combined with method factor II) but across classes every factor of one class occurs in combination with every factor in the other classes.

In specifying this, we are specifying the systematic structure of a set of measures. We are specifying the intended or designed factorial structure as it were. It will often be of interest to estimate the magnitude of each of these apriori factors in the measures that by design should contain them. In such cases, one will be interested in specifying loadings even where they are zero or minimal, as in learning that on self-description for "talkativeness" the method component is negligible. Ordinary passive, descriptive, free-floating, rotating factor analysis may only with awkwardness be used for this purpose. Usually it will not have extracted enough factors (e.g., one for each method and at least one for each trait); factors of generally low loadings are not identified; rotation of one oblique axis

---

[1] Supported by Project C 998, Contract 3-20-001, with the Educational Media Branch, Office of Education, U. S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

through each group of measures defining an apriori factor is not a readily available routine; estimation of number of factors and rotation uses up degrees of freedom in a manner irrelevant for this purpose; etc.

Thus a procedure that would go more directly to factors of apriori location could be a useful addition to our repertory of techniques. Such a procedure is explored in this paper.

It is convenient to start off with the small Vernon-Cronbach-Conrad matrix of Table 1. In the original presentation of the multitrait-multimethod matrix, it was discussed as follows:

> "Cronbach (1949, p. 277) and Vernon (1957, 1958) have both discussed the multitrait-multimethod matrix shown in Table [1], based upon data originally presented by H. S. Conrad. Using an approximative technique, Vernon estimates that 61% of the systematic variance is due to a general factor, that 21½% is due to the test-form factors specific to verbal or to pictorial forms of items, and that but 11½% is due to the content factors specific to electrical or to mechanical contents. Note that for the purposes of estimating validity, the interpretation of the general factor, which he estimates from the .49 and .45 heterotrait-heteromethod values, is equivocal. It could represent desired competence variance, representing components common to both electrical and mechanical skills -- perhaps resulting from general industrial shop experience, common ability components, overlapping learning situations, and the like. On the other hand, this general factor could represent overlapping method factors, and be due to the presence in both tests of multiple choice item format, IBM answer sheets, or the heterogeneity of the *S*s in conscientiousness, test-taking motivation, and test-taking sophistication. Until methods that are still more different and traits that are still more independent are introduced into the validation matrix, this general factor remains uninterpretable. From this standpoint it can be seen that 21½% is a very minimal estimate of the total test-form variance in the tests, as it represents only test-form components specific to the verbal or the pictorial items, i.e., test-form components which the two forms do *not* share. Similarly, and more hopefully, the 11½% content variance is a very minimal estimate of the total true trait variance of the tests, representing only the true trait variance which electrical and mechanical knowledge do *not* share" (Campbell & Fiske, 1959, pp. 92-93).

Table 1. Mechanical and Electrical Knowledge Measured by Verbal and Pictorial Items, after Vernon, Cronbach, and Conrad.

		Verbal Items		Pictorial Items	
		A_1	B_1	A_2	B_2
Verbal Items					
Mechanical Facts	A_1	(.89)	.63	.61	.49
Electrical Facts	B_1	.63	(.71)	.45	.51
Pictorial Items					
Mechanical Facts	A_2	.61	.45	(.82)	.64
Electrical Facts	B_2	.49	.51	.64	(.67)

According to the apriori model, the factorial composition of the four measures is as indicated in the top part of Table 2. The M_{1A_1}'s etc. are the factor loadings to be solved for, all blanks, i.e. all other possible factor loadings are assumed zero. In the second section of the table are illustrated the composition of the correlation coefficients. Remember that if there is a zero loading for a given factor in either of the measures, then that factor does not contribute to the correlation. In this part of Table 2, are indicated the products of the loadings of the two measures on that column's factor. The reliability coefficient ($r_{A_1A_1}$) is epitomized as the correlation between two measures of the same trait (A and G) measured by the same method (M_1). The correlation $r_{A_1B_1}$ has a similar composition except for the trait-A-specific factor. If the loadings of A_1 and B_1 on Method 1 were equal, then the difference between $r_{A_1A_1}$ and $r_{A_1B_1}$ would

Table 2. Apriori Factorial Composition of the Table 1 Measures.

		Method M_1 (Verbal)	Factors M_2 (Pictorial)	Trait T_A (Mechanical)	Factors T_B (Electrical)	General Factor G
Measure Composition	A_1	M_{1A_1}		$T_{A_{A_1}}$		G_{A_1}
	B_1	M_{1B_1}			$T_{B_{B_1}}$	G_{B_1}
	A_2		M_{2A_2}	$T_{A_{A_2}}$		G_{A_2}
	B_2		M_{2B_2}		$T_{B_{B_2}}$	G_{B_2}
Correlation Composition						
	$r_{A_1A_1} =$	$+ M_{1A_1} M_{1A_1}$		$+ T_{A_{A_1}} T_{A_{A_1}}$		$+ G_{A_1} G_{A_2}$
	$r_{A_1B_1} =$	$+ M_{1A_1} M_{1B_1}$				$+ G_{A_1} G_{B_1}$
	$r_{A_1A_2} =$			$+ T_{A_{A_1}} T_{A_{A_2}}$		$+ G_{A_1} G_{A_2}$
	$r_{A_1B_2} =$					$+ G_{A_1} G_{B_2}$
	$r_{B_1B_1} =$	$+ M_{1B_1} M_{1B_1}$			$+ T_{B_{B_1}} T_{B_{B_1}}$	$+ G_{B_1} G_{B_1}$
	$r_{B_1A_2} =$					$G_{B_1} G_{A_2}$
	$r_{B_1B_2} =$				$+ T_{B_{B_1}} T_{B_{B_2}}$	$+ G_{B_1} G_{B_2}$
	$r_{A_2A_2} =$		$+ M_{2A_2} M_{2A_2}$	$+ T_{A_{A_2}} T_{A_{A_2}}$		$+ G_{A_2} G_{A_2}$
	$r_{A_2B_2} =$		$+ M_{2A_2} M_{2B_2}$			$+ G_{A_2} G_{B_2}$
	$r_{B_2B_2} =$		$+ M_{2B_2} M_{2B_2}$		$+ T_{B_{B_2}} T_{B_{B_2}}$	$+ G_{B_2} G_{B_2}$

be interpretable as the contribution of the A_1 loading on Trait A to $r_{A_1A_1}$. While we know that they are not equal, the strategy of the present assumption is to assume that they are to get a first approximative solution. This solution will then be improved by an iterative process. To spell out the assumptions and procedure more thoroughly:

$$r_{A_1A_1} = M_{A_1}{}^2 + T_{A_{A_1}}{}^2 + G_{A_1}{}^2$$

$$r_{A_1B_1} = M_{A_1}M_{B_1} + G_{A_1}G_{B_1}$$

if $M_{A_1} = M_{B_1}$ and $G_{A_1} = G_{B_1}$

then $r_{A_1B_1} = M_{A_1}{}^2 + G_{A_1}{}^2$

and $r_{A_1A_1} - r_{A_1B_1} = T_{A_{A_1}}{}^2$

Another example:

$$r_{A_1A_2} = T_{A_{A_1}} T_{A_{A_2}} + G_{A_1} G_{A_2}$$

$$r_{A_1B_2} = G_{A_1} G_{B_1}$$

if $T_{A_{A_1}} = T_{A_{A_2}}$ and $G_{A_1} = G_{A_2} = G_{B_1}$

then $r_{A_1A_2} = T_{A_{A_1}}{}^2 + G_{A_1}{}^2$

$$r_{A_1B_2} = G_{A_1}{}^2$$

$$r_{A_1A_2} - r_{A_1B_2} = T_{A_{A_1}}{}^2$$

If we have usable reliabilities, then there are two such estimates for each factor loading, each a pair of correlations sharing one variable. The full set of first round estimates is shown in Table 3 (p. 7).

It may seem encouraging that with the exception of loadings estimated for measure A_1, the pairs of estimates of the same loading are in good agreement. This probably reflects more on the happenstance appropriateness of the transitional assumption of equal loadings in this specific instance than on the eventual goodness of fit of the model. It is more important to note that even though equal loadings have had to be assumed in this first round solution, the loadings do not come out equal. For example, in order for $\sqrt{r_{B_1B_1} - r_{B_1B_2}}$ to generate loadings of B_1 on Method 1, we have had to assume that the loadings of B_1 and B_2 on Trait B were the same. But the estimates of these loadings are consistently different. The assumption of equality has blurred and underestimated these differences, but has left a trace of them which the iterative procedure can amplify.

In this initial exploration, the several estimates for the same factor loading have been arithmetically averaged. Perhaps a geometric mean would be more appropriate. These averages, and the single values for the G factor loadings, based upon the square roots of $r_{A_1B_2}$ and $r_{A_2B_1}$ provide the basis for the iterative process.

The basic step in the iterative process is to assume that these first round estimates are correct for all but one of the factor loadings composing a given correlation, and to solve for that one. This partakes of that very general bootstrapping process in science in which one assumes the truth of the great bulk of current "knowledge" in a field, while

Table 3. First Round Estimates of Factor Loadings

	Method	Factors	Trait	Factors	
	1	2	A	B	G
A_1	$\sqrt{r_{A_1A_1} - r_{A_1A_2}} = .529$		$\sqrt{r_{A_1A_1} - r_{A_1B_1}} = .510$		$\sqrt{r_{A_1B_2}} = .700$
	$\sqrt{r_{A_1B_1} - r_{A_1B_2}} = .374$		$\sqrt{r_{A_1A_2} - r_{A_1B_2}} = .346$		
	Ave. = .452		Ave. = .428		
B_1	$\sqrt{r_{B_1B_1} - r_{B_1B_2}} = .447$			$\sqrt{r_{B_1B_1} - r_{B_1A_1}} = .283$	$\sqrt{r_{B_1A_2}} = .670$
	$\sqrt{r_{B_1A_1} - r_{B_1A_2}} = .424$			$\sqrt{r_{B_1B_2} - r_{B_1A_2}} = .245$	
	Ave. = .436			Ave. = .264	
A_2		$\sqrt{r_{A_2A_2} - r_{A_2A_1}} = .458$	$\sqrt{r_{A_2A_2} - r_{A_2B_2}} = .424$		$\sqrt{r_{A_2B_1}} = .670$
		$\sqrt{r_{A_2B_2} - r_{A_2B_1}} = .436$	$\sqrt{r_{A_2A_1} - r_{A_2B_1}} = .400$		
		Ave. = .447	Ave. = .412		
B_2		$\sqrt{r_{B_2B_2} - r_{B_2B_1}} = .400$		$\sqrt{r_{B_2B_2} - r_{B_2A_2}} = .173$	$\sqrt{r_{B_2A_1}} = .700$
		$\sqrt{r_{B_2A_2} - r_{B_2A_1}} = .387$		$\sqrt{r_{B_2B_1} - r_{B_2A_1}} = .141$	
		Ave. = .394		Ave. = .157	

critically examining and revising one portion of it (Campbell, 1966). This provides us with several new estimates for each factor loading. Since the G factor enters into each correlation, we get four estimates of the G loading for each measure, rather than the one estimate available for the first round. For the other loadings, it turns out that there are two estimates each. The computation of the Method 1 loading for A_1 from $r_{A_1 A_1}$ is as follows:

$$r_{A_1 A_1} = M_{1A_1}^2 + T_{A_{A_1}}^2 + G_{A_1}^2$$

$$.89 = M_{1A_1}^2 + .43^2 + .70^2$$

$$M_{1A_1}^2 = .89 - .18 - .49 = .22$$

$$M_{1A_1} = .47$$

The same loading is involved in $r_{A_1 B_1}$, which is solved as follows:

$$r_{A_1 B_1} = M_{1A_1} M_{1B_1} + G_{A_1} G_{B_1}$$

$$.63 = M_{1A_1} .44 + .70 \cdot .67$$

$$M_{1A_1} = \frac{.63 - .47}{.44} = .36$$

Taking each loading in turn, solving for it in all possible correlations, and averaging these new values gives us a set of new estimates. These could then be used as the tentatives in a second round of solutions. In practice, this has not been found to be effective. Such a full step seems to overcorrect for initial misestimates, leading to an oscillation in the iteration rather than a convergence. For this reason, a "half-step" has been employed, the round two estimates being made up of an average of these new values and the round one estimates. Such round two estimates are then used in generating round three, etc. Table 4 gives the round-by-round factor loadings employed, with the individual solution values from round one, upon which round two is based. The original correlations and the correlations predicted by each round's factor loadings are given in Table 5, along with the mean and ${}_{o}\sigma$ of residuals also indicated. (This ${}_{o}\sigma$ is a standard deviation of residuals from zero, rather than from the mean residual, because in this mode of analysis, the mean residual frequently departs from zero. So computing the ${}_{o}\sigma$ insures that it be a meaningful index of goodness of fit.)

In this particular case, the iterative procedure does not produce an entirely stable solution within twelve rounds, although in part this may be due to the relative excellence of the initial solutions. Note that the second round is distinctly worse than the first. From the fifth round on, the fit is slightly better than the first, but not much. On grounds of both mean residual and ${}_{o}\sigma$, the round seven is perhaps the best. For a justification of the half-step procedure (developed on matrices of another type to be discussed next) note that if the average solution values derived from round one were to be used in place of the present round 2, the fit

Table 4. Round-by-round factor loadings for each variable

	A_1			B_1			A_2			B_2		
	M_{1A_1}	$T_{A_{A_1}}$	G_{A_1}	M_{1B_1}	$T_{B_{B_1}}$	G_{B_1}	M_{2A_2}	$T_{A_{A_2}}$	G_{A_2}	M_{2B_2}	$T_{B_{B_2}}$	G_{B_2}
Round 1	.452	.428	.700	.436	.264	.670	.447	.412	.670	.394	.157	.700
Solutions	.466	.443	.709	.437	.266	.671	.448	.414	.671	.394	.158	.700
	.369	.342	.646	.456	.261	.619	.434	.329	.620	.383	.155	.700
			.648			.670			.663			.693
			.700			.670			.670			.700
Average Solution	.418	.393	.677	.397	.264	.658	.441	.372	.656	.389	.157	.698
Round 2	.435	.411	.689	.417	.264	.664	.444	.392	.663	.392	.157	.699
Round 3	.446	.442	.693	.420	.283	.668	.454	.396	.667	.395	.164	.702
Round 4	.443	.420	.689	.411	.278	.665	.453	.389	.665	.389	.155	.700
Round 5	.449	.425	.690	.412	.291	.667	.458	.389	.666	.390	.160	.702
Round 6	.449	.424	.689	.407	.289	.666	.458	.385	.665	.387	.156	.701
Round 7	.451	.427	.689	.406	.293	.667	.462	.385	.673	.387	.158	.702
Round 8	.451	.424	.687	.403	.292	.666	.457	.376	.669	.382	.157	.701
Round 9	.453	.428	.688	.403	.255	.667	.461	.378	.670	.383	.161	.702
Round 10	.454	.428	.688	.402	.293	.666	.461	.376	.674	.381	.160	.702
Round 11	.455	.426	.687	.402	.294	.666	.459	.371	.671	.378	.161	.702
Round 12	.457	.429	.688	.402	.295	.666	.463	.373	.672	.380	.164	.702

Table 5. Original and predicted correlations round by round, with mean and σ^1 of residuals

	$r_{A_1A_1}$	$r_{A_1B_1}$	$r_{A_1A_2}$	$r_{A_1B_2}$	$r_{B_1B_1}$	$r_{B_1A_2}$	$r_{B_1B_2}$	$r_{A_2A_2}$	$r_{A_2B_2}$	$r_{B_2B_2}$	Mean Res.	σ_{res}
Original	.890	.630	.610	.490	.710	.450	.510	.820	.640	.670		
Round 1	.877	.666	.645	.490	.709	.449	.510	.818	.645	.670	-.0059	.0165
(Full-Step)	(.788)	(.611)	(.620)	(.473)	(.660)	(.432)	(.501)	(.763)	(.629)	(.663)	(.028)	(.0419)
Round 2	.833	.639	.618	.482	.684	.440	.506	.790	.637	.667	.0124	.0228
Round 3	.857	.650	.629	.486	.703	.446	.515	.808	.648	.676	.0002	.0148
Round 4	.847	.640	.622	.482	.688	.442	.509	.799	.642	.665	.0084	.0179
Round 5	.858	.645	.625	.484	.699	.444	.515	.805	.646	.671	.0028	.0141
Round 6	.856	.642	.621	.483	.693	.443	.512	.800	.643	.666	.0061	.0148
Round 7	.860	.643	.628	.484	.696	.449	.515	.815	.651	.668	.0011	.0134
Round 8	.855	.639	.619	.482	.691	.446	.513	.798	.644	.662	.0071	.0155
Round 9	.862	.641	.623	.483	.694	.447	.516	.804	.647	.665	.0038	.0134
Round 10	.863	.641	.625	.483	.691	.449	.514	.808	.649	.665	.0032	.0134
Round 11	.861	.640	.619	.482	.692	.447	.514	.799	.644	.661	.0061	.0141
Round 12	.866	.642	.622	.483	.692	.448	.516	.805	.648	.664	.0024	.0134

would be much worse than any of those in Table 5, providing a mean residual of .028 and a $_{o}\sigma$ of .0419, almost twice as large as the values of the half-step round two. (These values are indicated as the "Full Step" entries of Table 5.)

The computations provided in this paper have all been done by hand, and only spot-checked for accuracy. The deadlines for this report have prevented preparation of a computer program and more extensive iteration. It remains an open question whether or not further iteration would produce a better and stable solution. It can be noted that while all variables have oscillated in direction of shift, the best rounds do not represent a return to the same loading pattern. Round 12 is not a return to round 7, for example. Further, some loadings are at round 12 at their most extreme point so far (M_{1A_1}, M_{2A_2}, and $T_{A_{A_2}}$) and may be started on a trend of development that would provide further progress. But by normal factor analytic standards, the final fit of round 12 is quite good.

Usable reliabilities are not always present. While this means fewer bases for estimating the factor loadings, at least one for each is still available, and the iterative procedure can again be used. Tables 6 and 7 parallel Tables 4 and 5, with the exception that the four reliabilities ($r_{A_1A_1}$ etc.) are not employed.

The final factor loadings are, of course, different from the previous set, due to the removed influence of the four reliabilities or due to only 6/10 of the previous empirical base being employed. The better fit, i.e.,

Table 6. Procedures of Tables 3 and 4 repeated without using reliability coefficients

	A_1			B_1			A_2			B_2		
	M_{1A_1}	$T_{A_{A_1}}$	G_{A_1}	M_{1B_1}	$T_{B_{B_1}}$	G_{B_1}	M_{2A_2}	$T_{A_{A_2}}$	G_{A_2}	M_{2B_2}	$T_{B_{B_2}}$	G_{B_2}
Round 1	.374	.346	.700	.424	.245	.670	.436	.400	.670	.387	.141	.700
2	.377	.350	.701	.427	.268	.673	.439	.404	.672	.390	.154	.702
3	.374	.347	.700	.399	.258	.671	.435	.401	.671	.387	.148	.700
4	.388	.348	.703	.414	.264	.674	.437	.402	.672	.389	.152	.701
5	.383	.345	.701	.408	.257	.682	.436	.398	.671	.388	.148	.700
6	.378	.349	.701	.403	.240	.678	.437	.402	.670	.389	.138	.699
7	.382	.349	.702	.407	.251	.679	.440	.402	.670	.392	.144	.700
8	.379	.349	.701	.404	.247	.677	.438	.402	.669	.391	.142	.699
9	.382	.350	.702	.407	.254	.678	.439	.403	.669	.392	.146	.700
10	.380	.349	.702	.405	.247	.677	.439	.402	.669	.393	.142	.700
11	.382	.349	.702	.407	.251	.677	.439	.402	.668	.393	.144	.700
12	.382	.350	.702	.407	.251	.677	.439	.403	.668	.393	.144	.700
Table 4 Rnd. 12	.457	.429	.688	.402	.295	.666	.463	.373	.672	.380	.164	.702

Table 7. Correlations predicted by Table 6, with M_{res} and σ_{res}

	$r_{A_1B_1}$	$r_{A_1A_2}$	$r_{A_1B_2}$	$r_{B_1A_2}$	$r_{B_1B_2}$	$r_{A_2B_2}$	Mean Res.	σ_{res}
Original	.630	.610	.490	.450	.510	.640		
Round 1	.628	.607	.490	.449	.504	.638	.0023	.0030
2	.633	.612	.492	.452	.514	.643	-.0027	.0028
3	.619	.609	.490	.450	.508	.638	.0027	.0164
4	.634	.612	.493	.453	.513	.641	-.0027	.0028
5	.634	.608	.491	.458	.515	.639	-.0025	.0042
6	.628	.610	.490	.454	.507	.638	.0005	.0023
7	.632	.611	.491	.455	.511	.641	-.0018	.0023
8	.628	.609	.490	.453	.508	.639	.0005	.0018
9	.631	.611	.491	.454	.512	.640	-.0015	.0020
10	.629	.610	.491	.453	.509	.641	-.0005	.0014
11	.631	.609	.491	.452	.510	.640	-.0005	.0011
12	.631	.610	.491	.452	.510	.640	-.0007	.0008
From Table 5, Round 12	.642	.622	.483	.448	.516	.648	.0024	.0134

the smaller $_o\sigma$ of residuals, is due to the smaller number of values being fitted, the smaller number of degrees of freedom for the same number of factors. The process of convergence is much more marked, although again, the first solution is better than some of the intermediate rounds.

As a second example, let us take a "response set" study. Table 8 shows a matrix of correlations among four scales measuring "the same" attitude. Two are worded in the same direction (e.g., pro) and two in the opposite (con). One pro and one con are administered on the same occasion, one each a week later, but this occasions factor is disregarded in this analysis.

Table 8

	P_1	C_1	P_2	C_2
P_1		.62	.78	.40
C_1	.62		.50	.68
P_1	.78	.50		.33
C_2	.40	.68	.33	

In this instance, all variables have apriori factor loadings on both of two factors. The first is a trait factor, on which all have positive loadings, the second is a direction of wording factor, for which the con items have loadings of opposite sign from the pro.

Table 9 shows the apriori factor model for each variable and each correlation. The first estimations of factor loadings are made as follows:

$$r_{P_1P_2} = +M_{P_1}M_{P_2} + T_{P_1}T_{P_2}$$

$$r_{P_1C_1} = -M_{P_1}M_{C_1} + T_{P_1}T_{C_1}$$

if $M_{P_1} = M_{P_2} = M_{C_1}$ and $T_{P_1} = T_{P_2} = T_{C_1}$, then

$$r_{P_1P_2} = +M_{P_1}^2 + T_{P_1}^2$$

$$r_{P_1C_1} = -M_{P_1}^2 + T_{P_1}^2$$

$$r_{P_1P_2} - r_{P_1C_1} = 2M_{P_1}^2$$

$$M_{P_1} = \sqrt{\frac{r_{P_1P_2} - r_{P_1C_1}}{2}}$$

$$r_{P_1P_2} + r_{P_1C_1} = 2T_{P_1}^2$$

$$T_{P_1} = \sqrt{\frac{r_{P_1P_2} + r_{P_1C_1}}{2}}$$

Table 9. Apriori factor model for each variable and correlation

	Method	Trait
P_1	$+ M_{P_1}$	$+ T_{P_1}$
C_1	$- M_{C_1}$	$+ T_{C_1}$
P_2	$+ M_{P_2}$	$+ T_{P_2}$
C_2	$- M_{C_2}$	$+ T_{C_2}$

$$r_{P_1 C_1} = - M_{P_1} M_{C_1} + T_{P_1} T_{C_1}$$

$$r_{P_1 P_2} = + M_{P_1} M_{P_2} + T_{P_1} T_{P_2}$$

$$r_{P_1 C_2} = - M_{P_1} M_{C_2} + T_{P_1} T_{C_2}$$

$$r_{C_1 P_2} = - M_{C_1} M_{P_2} + T_{C_1} T_{P_2}$$

$$r_{C_1 C_2} = + M_{C_1} M_{C_2} + T_{C_1} T_{C_2}$$

$$r_{P_2 C_2} = - M_{P_2} M_{C_2} + T_{P_2} T_{C_2}$$

Two such estimates are available for each measure, involving, however, the same larger (e.g., $r_{P_1 P_2}$ or $r_{C_1 C_2}$) coefficient. In best practice, the two would probably be computed and then averaged. However, in the illustrative computational model here provided, the two differences have been averaged and the square root of this average employed as the first estimate of the method factor. In each instance, the inference base is limited to comparisons among correlations all of which involve the measure in question.

Table 10 shows the first round estimates of the factor loadings. These have been used, as in the previous example, to solve each coefficient for one loading, assuming the others known. In this case, estimates of each factor come from each of the three coefficients involving the measure. These three are averaged, and the next round represents a half-step move toward this average. Table 11 shows the predicted correlations for each round. Round two shows a definite gain over round one, and a slight trend toward improvement continues to round four, when there is a slight recession in σ_{res}.

The final illustration is still more incomplete. It explores the problems of direct factoring for larger matrices, where the problem of optimal approach is not fully solved. In Table 12 is shown a hypothetical multitrait multimethod matrix, which should be solvable without residual, being generated from a set of loadings conforming exactly to the hypothesized pattern. These loadings are shown in Table 13.

Table 10. Iterative solutions of factor loadings for the direction-of-wording illustration

	Method Factor				Trait Factor			
	M_{P_1}	M_{C_1}	M_{P_2}	M_{C_2}	T_{P_1}	T_{C_1}	T_{P_2}	T_{C_2}
Round 1	.367	-.245	.428	-.397	.803	.787	.773	.723
Solutions	.049	-.033	.434	-.492	.902	.884	.776	.680
	.372	-.253	.442	-.453	.806	.782	.769	.740
	.455	-.280	.577	-.535	.755	.806	.691	.647
Average	.292	-.189	.484	-.493	.821	.824	.745	.689
Round 2	.330	-.217	.456	-.445	.812	.806	.759	.706
Solutions	.159	-.104	.496	-.525	.858	.852	.775	.673
	.359	-.245	.515	-.511	.829	.789	.743	.724
	.389	-.249	.463	-.451	.775	.826	.755	.702
Average	.302	-.199	.491	-.496	.821	.822	.758	.700
Round 3	.316	-.208	.474	-.471	.817	.814	.759	.703
Solutions	.217	-.143	.506	-.552	.842	.839	.771	.672
	.337	-.249	.566	-.518	.830	.789	.735	.715
	.370	-.229	.432	-.429	.781	.828	.787	.729
Average	.308	-.207	.501	-.500	.818	.819	.764	.705
Round 4	.312	-.208	.488	-.486	.818	.817	.762	.704
Solutions	.232	-.155	.502	-.564	.838	.837	.767	.674
	.321	-.251	.589	-.504	.824	.789	.736	.709
	.362	-.216	.425	-.423	.784	.822	.806	.744
Average	.305	-.207	.505	-.497	.815	.816	.770	.709
Round 5	.309	-.208	.497	-.492	.817	.817	.766	.707

Table 11. Original and predicted correlations iteratively

	$r_{P_1 C_1}$	$r_{P_1 P_2}$	$r_{P_1 C_2}$	$r_{C_1 P_2}$	$r_{C_1 C_2}$	$r_{P_2 C_2}$	$M_{res.}$	$_o\sigma_{res.}$
Original	.62	.78	.40	.50	.68	.33		
Round 1	.54	.78	.43	.50	.67	.39	.0000	.042
2	.58	.77	.43	.51	.67	.33	.0033	.022
3	.60	.77	.43	.52	.67	.31	.0017	.020
4	.60	.78	.42	.52	.68	.30	.0017	.019
5	.60	.78	.43	.52	.68	.30	.0000	.021

Table 12. The "True" Correlation Matrix generated by the factor loadings of Table 13, and set as the problem for the iterative factoring shown in Tables 14, 15 and 16

	A_1	B_1	C_1	D_1	E_1	A_2	B_2	C_2	D_2	E_2
A_1	(.33)	.23	.15	.04	.06	.35	.12	.12	.00	.00
B_1	.23	(.59)	.21	.08	.12	.18	.35	.12	.00	.00
C_1	.15	.21	(.50)	.14	.09	.09	.06	.40	.12	.00
D_1	.04	.08	.14	(.17)	.06	.00	.00	.06	.21	.00
E_1	.06	.12	.09	.06	(.34)	.00	.00	.00	.00	.15
A_2	.35	.18	.09	.00	.00	(.68)	.35	.37	.30	.25
B_2	.12	.35	.06	.00	.00	.35	(.45)	.28	.24	.20
C_2	.12	.12	.40	.06	.00	.37	.28	(.66)	.39	.25
D_2	.00	.00	.12	.21	.00	.30	.24	.39	(.70)	.30
E_2	.00	.00	.00	.00	.15	.25	.20	.25	.30	(.34)

Table 13. "True" factor loadings for Table 12

	Method Factors		Trait Factors							
	M_1	M_2	T_A	T_B	T_C	T_D	T_E	T_{AB}	T_{ABC}	T_{CD}
A_1	.20		.40					.20	.30	
B_1	.40			.50				.30	.30	
C_1	.30				.40				.30	.40
D_1	.20					.30				.20
E_1	.30						.50			
A_2		.50	.50					.30	.30	
B_2		.40		.40				.30	.20	
C_2		.50			.40				.40	.30
D_2		.60				.50				.30
E_2		.50					.30			

When one moves away from the two traits by two methods, or four variable matrix to something larger, as this five traits by two methods, one both gains and loses. For the method 1 loading in A_1, we now have four or five correlations on which to base estimates. (Five if a usable reliability is present.) But on a pure apriori basis we now have increased the number of trait factors much more than the number of measures. If we do not assume that traits are independent, then on the basis of examining all of the apriori factors we should examine A, B, C, D, E; AB, AC, AD, AE, BC, BD, BE, CD, CE, DE (i.e., factors shared by pairs but not otherwise); ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE; and ABCD, ACDE, ABDE, ABCE, BCDE; or 30 factors plus the G factor unallocatable between a M_{1-2} factor shared by methods, and an ABCDE factor shared by all traits.

In the face of this apriori complexity, the precise estimation of trait factors has been given up, the focus instead being on getting for each measure an estimate of its method component, (more accurately, its method-specific component), and of its "trait communality," a not specifically allocated pool of trait factors. (Ordinary factoring procedures could no doubt be modified for estimating trait factors, using the apriori communality estimates of the present system.)

The procedure employed in each round has been to treat the 10 x 10 matrix as ten 4 x 4 matrices, setting up one for each pairing of traits, AB, AC, etc. Thus the submatrix A_1, B_1, A_2, B_2, has been factored in the manner of Tables 2, 3, 4, and 5; so has the partially overlapping Table A_1, C_1, A_2, C_2; and A_1, D_1, A_2, D_2, etc. In this procedure, for the AB run the apriori trait factors have been pooled as follows:

with A go AC, AD, AE, ACD, ACE, ADE, ACDE

with B go BC, BD, BE, BCD, BCE, BDE, BCDE

with G go AB, ABC, ABD, ABE, ABCD, ABCE, ABDE (and ABCDE and M_{1-2}),

not contributing are C, D, C, CD, CE, DE, and CDE. In these ten 4x4 studies, the resultant trait factors are not comparable. The $T_{A_{A_1}}$ of A_1, B_1, A_2, B_2 is not the same construct as the $T_{A_{A_1}}$ of A_1, C_1, A_2, C_2 (as shown in the "true" values of Table 14). But the M_{1A_1}'s are the same construct. The iterative process has as a result pooled all the method loading estimates of the ten 4 x 4 matrices to provide the basis for the next round. Thus in the iteration all ten matrices are moved one round, pooling of methods estimates is done across all ten, trait loadings estimates are done separately, and a next round begun. Again, inspite of the much broader base of the estimates, the next round method loadings are moved only half of the distance to the mean estimate. The particular matrix of Table 12 requires making another procedural rule: from the nature of the case, negative loadings on methods factors are not allowed. (While they are allowed in the second illustration, Tables 8, 9, and 10, this is determined apriori by the test-scoring formula, and would not be allowed to change sign to improve a factorial fit.) Due to the pattern of loadings chosen, in Table 12 the reliability of A_1 is paradoxically lower than its validity. This provides a first round estimate of the M_{1A_1} which would be $\sqrt{.33 - .35}$. In such cases, a value of zero is employed instead.

Serious clerical errors discovered in the work on this problem prevent us from presenting tables of the results, but the approach seems feasible. For the M factors, there are five estimates for each, one based on the reliability-validity difference and four based on the $r_{A_1B_1} - r_{A_1B_2}$ type comparison. These are averaged and that value used in all four of the ten 4 x 4 studies in which that variable appears.

While no poolable factor estimates emerge for the trait factors, the total trait communality for each measure should be the same in each of the four 4 x 4 matrices in which it participates. This trait communality is the sum of the squares of the trait loadings and the G loading. This suggests that in instances containing sampling error, at least, the four estimates of this communality be averaged, and then the loadings corrected so that they produce this average communality. Table 14 illustrates this process. On the basis of procedures such as in Tables 2, 3, and 4, the "Round 2" loadings on $T_{A_{A_1}}$ and G_{A_1} are estimated. The "true" values of these loadings are also indicated, computed from Table 13 by combining factors which are pooled in the 4 x 4 submatrices. Thus in the AB study, the .400 value for $T_{A_{A_1}}$ comes straight from the A column of Table 11 because there are no trait factors other than T_A which A_1 fails to share with B_1. The G factor in the AB study pools the loadings of Table 13 in the T_{AB} and T_{ABC} columns (e.g., $G_{A_1} = \sqrt{T_{AB_{A_1}}{}^2 + T_{ABC_{A_1}}{}^2} = \sqrt{.04 + .09} = .360$). The "true" $T_{A_{A_1}}$ differs in the AC study because in Table 13 A_1 has both $T_{A_{A_1}}$ and $T_{AB_{A_1}}$ not shared with C_1, etc.

Table 14. Correcting trait-loading estimates for average trait communality. Illustrated for measure A_1 .

		$T_{A_{A_1}}$	G_{A_1}	$T^2{}_{A_{A_1}}$	$G^2{}_{A_1}$	h^2
AB	Round 1	.398	.346	.158	.120	.278
	Corrected	.419	.364	.175	.133	.308
	True	.400	.360	.160	.130	.290
AC	Round 1	.452	.346	.204	.120	.324
	Corrected	.441	.338	.194	.114	.308
	True	.447	.300	.200	.090	.290
AD	Round 1	.566	.000	.320	.000	.320
	Corrected	.555	.000	.308	.000	.308
	True	.538	.000	.290	.000	.290
AE	Round 1	.556	.000	.309	.000	.309
	Corrected	.555	.000	.308	.000	.308
	True	.538	.000	.290	.000	.290

In Table 14, the trait communality estimates are computed as the sum of the squares of the $T_{A_{A_1}}$ and G_{A_1} loadings. These communality values should all be the same, as note the "true" values are. As a best estimate, we use the arithmetic average of them, which is .308, and then magnify or reduce the squared factor loadings so that in the "corrected" values they have this average, and from these squares the corrected loadings are taken. These corrected values are, on the average, closer to the true values than the simple round one values. (Perhaps some approach could be worked out which would work only with method factors and the pooled trait communality or its square root, h_t, obviating the approach of breaking the matrix up into four-by-four studies.)

In the end we can state for each measure its theoretical reliability, i.e., the communality or sum of the squares of its final loadings, and break this into the methods component (the square of its method loading) and its trait component (the trait communality), all values broadly based upon a pooling of several sources of estimates. While such results will be specific to the matrix of study, and ambiguous where none of the heterotrait-heteromethod correlations get down to a true zero, such indices should none the less be a considerable addition to the evaluational indices presently available.

Discussion

Number of factors relative to number of variables. The procedures here described certainly violate the sensible rule of factor analysis that the number of factors extracted be one (or even two) less than the number of measures correlated. Technically, of course, this refers to the number of *common* factors. If, as here, the reliabilities are taken as communalities, as empirical values to be employed in the factoring, then they provide the basis of estimating loadings on one *specific* factor for each variable, and we are already up to K + (K - 1 or 2) factors, where K is the number of measures. In terms of degrees of freedom, perhaps (K - 1)(K - 1) describes the situation better. But this is the rule used in cross-classifications containing K^2 independent cells. In our case if reliabilities are present we have but $\frac{K\,(K + 1)}{2}$ independent observations and if reliabilities are absent, but $\frac{K\,(K - 1)}{2}$. Perhaps the degrees of freedom would be $\frac{(K - 1)(K + 1 - 1)}{2}$, and with the reliabilities absent, $\frac{(K - 1)(K - 1 - 1)}{2}$. If so, in the case of the two-methods-two-traits study of Tables 2, 3, and 4 we had 6 degrees of freedom, and estimated twelve loadings spread over five factors. (The fact that not every factor has loadings on every variable may alleviate the problem somewhat.) In the same study re-run without reliabilities, we would have estimated the same twelve loadings spread over five factors, but with perhaps only three degrees of freedom? Or should we call these nine and five degrees of freedom? In any event, in the latter instance we certainly have no d. f.

left to test the goodness of fit, but at a descriptive level, we have estimated the twelve values in a way that seems to me to be useful and to reflect the data.

Size of matrix and accuracy of estimation. It is probably generally true of any factor analysis that the stability or accuracy of estimating a given factor loading or a given measure is greater the larger the size of the matrix, holding constant the size of the sample upon which correlations are computed. It is surmised that this would be particularly true of the present type. What is called for is not only demonstrations of the ability of the iterative process to reconstruct "perfect" cases, but also its ability to reconstruct these when "sampled" with small numbers of cases from models applied to normal random number computer simulations. In such a study, it would also be of interest to compare the operation of ordinary modes of factor analysis represented in computer programs.

Criteria of iteration. The present study has used the simple additive criteria of centroid factor analysis. It would seem a complex but not insuperable task to modify it to a Lawley (1940) maximum likelihood procedure whereby the goodness of fit to originally high correlations was weighted higher than that to low values. An adaptation of the criteria of Kaiser's alpha factor analysis (Kaiser & Caffrey, 1965) would also seem possible.

Other apriori approaches to factoring and/or summary statistics for multitrait-multimethod-matrices. A number of other approaches to the problem here addressed are available. As far as is known, the present paper does not duplicate any of these proposals, but the present author's knowledge of mathematics, analysis of variance and factor analysis does not rule out possible duplication. In any event, a high priority agendum is to relate these various approaches.

Working within the factor analytic tradition are Wolins (1965), Tucker (1958; 1963), Bechtoldt (1958; 1961), Horst (1959), Mosier (1939), Cliff (1962), Gibson (1963), and Madansky (1964). Working in the analysis of variance tradition are Stanley (1961; Stanley and Wiley, 1962) and Bock (1960; 1963).

Appropriateness of the additive model. It is a source of intellectual pain that this seemingly satisfactory solution to factoring multitrait-multimethod matrices comes at the same time as evidence of the inappropriateness of the factor analytic model to such matrices. (Campbell and O'Connell, 1965). When numerous empirical studies are examined, an interaction of the size of the trait and size of method factors is regularly found: the larger the shared trait factors between two traits, the larger the loading on the method factor. A given variable's method factor loading seems higher in a correlation with a similar trait than in a correlation with a very dissimilar trait. The method factor seems to operate multiplicatively, rather than additively. Each measure would be conceived of as having a method factor "loading" larger than one, which multiplied the shared trait covariance involved in a correlation. This multiplier can be estimated as the individual variable's "slope" when the heterotrait-monomethod values are plotted against their corresponding heterotrait-heteromethod variables. This slope may also turn out to be like $1/\sqrt{r_{A_1 A_2}}$ or a correction for attenuation process. That is, one can conceptualize the difference in method as attenuating the monomethod estimate of the intertrait correlation. Such initial estimates either from slope coefficients, as from O'Connell's computer program (Campbell & O'Connell, 1965), or from the reciprocal of the square root of the validity values, could then be refined by an iterative solution formula analogous to the one reported here. The formula might be regarded as:

$$r_{A_1 B_1} = M_{1A_1} \; M_{1B_1} \; (T_{AB_{A_1}} \; T_{AB_{B_1}} + G_{A_1} \; G_{B_1})$$

if one keeps the additive model for the trait variance. If $T_{AB_{A_1}} \; T_{AB_{B_1}} + G_{A_1} \; G_{B_1}$ equals $T_{AB_{A_1}} \; T_{AB_{B_2}} + G_{A_1} \; G_{B_2}$, then $r_{A_1 B_2} = (T_{AB_{A_1}} \; T_{AB_{B_1}} + G_{A_1} \; G_{B_1})$ and $r_{A_1 B_1} = M_{1A_1} \; M_{1B_1} \; r_{A_1 B_2}$, which can be conceived as $r_{A_1 B_1}$ equaling $r_{A_1 B_2}$ corrected for attenuation, if M_{1A_1} is construed, as suggested above, as analogous to $1/\sqrt{r_{A_1 A_2}}$. In either understanding, grounds for an iterative improvement of estimates seem present. The attenuation model would seem to rule out that gain in precision which comes from recognizing that B_1 and B_2 (etc.) have different loadings on T_B, T_{AB}, G, etc.

Summary

In multitrait multimethod matrices and other matrices with crossed factor classes, "direct" solutions for the apriori method and trait factors can be made by assuming that the two loadings involved in each factor product making up a correlation are equal. The average solution of each factor loading can then be entered into the apriori formula for each correlation for all loadings but one, which one is then solved for as an unknown. Each loading can be so solved for in turn, generating a new set of estimates. Moving the next round base half way toward the average of such solutions, an iterative process can be established which in many instances will converge. This approach to apriori factoring is judged worthy of further elaboration and trial by use.

References

Bechtoldt, H. P. Statistical tests of hypotheses in confirmatory factor analysis. Talk given to APA, September, 1958, Washington, D.C.

Bechtoldt, H. P. An empirical study of the factor analysis stability hypothesis. Psychometrika, 1961, 26, 405-432.

Bock, R. D. Components of variance analysis as a structural and discriminal analysis for psychological tests. British Journal of Statistical Psychology, 1960, 13, 151 ff.

Bock, R. D. Multivariate analysis of variance of repeated measurements. In C. W. Harris, Ed., Problems in measuring change. Madison: University of Wisconsin Press, 1963, pp. 85-103.

Campbell, D. T. Pattern matching as an essential in distal knowing. In K. R. Hammond, Ed., Probabilistic functionalism: the psychology of Egon Brunswik. New York: Holt, Rinehart and Winston, late 1965 or early 1966, in press.

Campbell, D. T. and Fiske, D. W., Convergent and discriminant validitation by the multitrait-multimethod matrix. Psychological bulletin, 1959, 56, 81-105.

Campbell, D. T. and O. Connell, E. J. Method factors in multitrait-multimethod matrices: multiplicative rather than additive? Mimeographed research report, Department of Psychology, Northwestern University, 1965.

Cliff, N. Analytic rotation to a functional relationship. Psychometrika, 1962, 27, 283-295.

Cronbach, L. J. Essentials of psychological testing. New York: Harper, 1949.

Gibson, W. A. On the symmetric treatment of an asymmetric approach to factor analysis. Psychometrika, 1963, 28, 423-426.

Horst, P. Relations among m sets of measures. Psychometrika, 1965, 30, 1-14.

Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh, 1940, 60, 64-82.

Madansky, A. Instrumental variables in factor analysis. Psychometrika, 1964, 29, 105-113.

Mosier, Charles. Determining a simple structure when loadings for certain tests are known." Psychometrika, 1939, 4, 149-162.

Stanley, J. C. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. Psychometrika 1961, 26, 205-220.

Stanley, J. C. and Wiley, D. E. Development and analysis of experimental design for ratings. Duplicated research report to the U. S. Office of Education, Department of Educational Psychology, University of Wisconsin, Madison, 1962.

Tucker, R. An inter-battery method of factor analysis. Psychometrika, 1958, Vol. 23, 111-136.

Tucker, L. R. Implications of factor analysis of three-way matrices for measurement of change. In C. W. Harris, Ed. Problems in measuring change. Madison: University of Wisconsin Press, 1963, pp. 122-137.

Vernon, P. E. Educational ability and psychological factors. Address to the Joint Education - Psychology Colloquium, University of Illinois, March 29, 1957.

Vernon, P. E. Educational testing and test-form factors. Princeton: Educational Testing Service, 1958. (Res. Bull. RB-58-3).

Wolins, L. A procedure for estimating the amount of trait, method, and error variance attributable to a measure. Duplicated research report on Project C998, Contract 3-30-001, Educational Media Branch, U. S. Office of Education.

30 September 1965

The Sense of Identity:

Measurement and an Examination of Correlates[1]

Frank Baker[2]

Northwestern University

---

[1] Supported in part by U.S. Office of Education Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

[2] Now at Lehigh University

Table of Contents

Page

List of Tables ii

Introduction 1

Correlates of a Sense of Identity. 8

  A Sense of Identity and a Sense of Trust. 8
  A Sense of Identity and a Sense of Industry 9
  A Sense of Identity and a Sense of Intimacy 10
  A Sense of Identity and Authoritarianism. 12

Method 13

  Construction of Scales. 13
  Sentence-Completion Instruments 14
  Subjects. 16
  Data-Processing Procedures. 16

Results and Discussion 18

  Reliability 18
  Relations Between Aspects of Identity 20
  Validity of Trust, Industry and Intimacy Measures 24
  Identity and Trust. 25
  Identity and Industry 25
  Identity and Intimacy 26
  Identity and Authoritarianism 27

Summary and Conclusions. 29

References 31

Appendix A 33

Appendix B 37

Appendix C 40

Appendix D 43

Appendix E 46

Appendix F 105

Tables

Page

1. Kuder-Richardson Reliability Coefficients for Scales and Sentence-Completion Measures. 19

2. Multitrait-Multimethod Matrix Based on Direct-Attitude and Sentence-Completion Measures of Identity Variables and Predicted Correlates. 21

3. Correlations Between Authoritarianism and Aspects of Identity. 28

The Sense of Identity:
Measurement and an Examination of Correlates

Frank Baker
Northwestern University

Erik Erikson (1950) has asserted that today the search for identity is as important as sexuality was in Freud's time, and has made the concept of ego identity a central variable in his explication of an epigenetic theory of the development of the personality (Erikson 1950a; 1950b; 1958; 1959). Erikson's theory grows out of the Freudian tradition and Erikson's own anthropological fieldwork. As Ruth Monroe (1960) has noted, Erikson shows how the classic Freudian genetic scheme may be extended from its emphasis on infantile sexuality and broadened to relate to the development of deep personal and social attitudes.

Erikson examines the growth of the personality in terms of a series of eight successive stages, "predetermined in the human organism's readiness to be aware of, and to interact with a widening social radius" (Erikson, 1959, p. 52). As the developing individual encounters different aspects of the social environment, each step becomes a potential crisis because of the attendant radical change in perspective. The problems of each stage can be solved in one of two polar directions which lead to the development of a series of alternative basic senses or attitudes: (1) trust versus mistrust, (2) autonomy versus shame and doubt, (3) initiative versus guilt, (4) industry versus inferiority, (5) identity versus identity diffusion, (6) intimacy versus isolation, (7) generativity versus self-absorption, (8) integrity versus disgust and despair. The psychosocial quality of each basic attitude becomes more differentiated, as the ego comes into the possession of a more intensive apparatus, even as society challenges and guides such extensions.

Adolescence is the stage at which Erikson postulates a crisis of identity. The term ego identity is used by Erikson to demote certain comprehensive gains which the individual, by the end of adolescence, must have derived from all pre-adult experience in order to be ready for the tasks of adulthood. The sense of ego identity has grown throughout childhood, but the rapid physiological changes that come with puberty force the young person to "refight many of the earlier battles." The developmental task is to integrate childhood identifications "with the basic biological drives, native endowment, and the opportunities offered in social roles." The alternative to the establishment of a sense of identity is the development of a sense of identity diffusion. While identity diffusion is temporarily unavoidable in this period of physical and psychological upheaval, the danger is that there may result a permanent inability to "take hold."

While Erikson asserts that "a sense of identity . . . can be defined and evidences for the presence of a dominant attitude of this kind can be described behavioristically" (Erikson, 1950b, p.63), there are relatively few empirical studies of this variable in the psychological literature.

Yufit conducted one of the first empirical studies which employed Erikson's concepts. Yufit's research focused primarily on Erikson's sixth stage of development in which the nuclear conflict is called Intimacy versus Isolation. His main conclusion was that Intimates and Isolates exist as gross personality types and that, "Each type is represented by a unique syndrome of behaviors on a descriptive level and a supporting set of dynamics on a more covert level" (Yufit, 1956, p. 26). Yufit also found that persons classified as Intimates had a better sense of identity than the Isolate group. The study utilized ratings based on interviews and projective data (Thematic Apperception Test) as its measures of intimacy versus isolation and of identity.

Recently in the Journal of Projective Techniques, Holt (1960) has encouraged the general use of Erikson's concept of ego identity in the interpretation of diagnostic tests such as the TAT as illustrated in this quote: "For its relevance to the issue of personal strengths as well as its implication for pathology, Erikson's concept of ego identity is one the diagnostic tester will do well to study closely and add to his stock of everyday tools" (Holt, 1960, p. 264).

In analyzing projective tests as a part of a larger study of culture change among the Bartok of Sumatra, Ross (1962) found one thematic picture in a specifically prepared picture story test based on the TAT elicited stories from six individuals which reflected a sense of not belonging. Apparently following Holt's advice, Ross opportunistically interpreted this to indicate that the subjects had a disturbed sense of identity.

In the first attempt to obtain a more objective measure of ego identity than ratings based on interviews or projective data, Gruen (1960) experimented with the real-ideal Q-sort discrepancy correlation score developed by Butler and Haigh (1954). However, it is doubtful that this is an appropriate measure of the construct of ego identity. The real-ideal Q-sort discrepancy score has often been defined as a measure of self-esteem, and Erikson (1950b) has made it a point to try to distinguish his concept of ego identity from the orthodox one of self-esteem. French and Kahn (1962) have amplified the distinction between self-esteem and self-identity: "Self-esteem may be defined as the average evaluation of the attributes of the self-identity, where each attribute is weighted according to its centrality. Another measure of self-esteem may be derived from discrepancies between the person's perceived attributes and the attributes of his ideal self, where the ideal self is conceived as the most desirable positions on the dimension of self-identity" (1962, p. 21).

This brief review of the literature indicates the scarcity of objective empirical studies of Erikson's concept of identity. Perhaps the main explanation is to be found in the lack of clarity in Erikson's own conceptualization of the term. He chooses to use his term identity with a number of connotations:

> At one time, then it will appear to refer to a conscious sense of individual identity; at another to an unconscious striving for a continuity of personal character; at a third, as a criterion for the silent doings of ego synthesis; and finally, as a maintenance of an inner solidarity with a group's ideals and identity (Erikson, 1959, p. 102).

In the introduction to a collection of papers by Erikson on identity, it is noted, "Erikson's theory (like much of Freud's) ranges over phenomenological, specifically clinical psychoanalytic and general psychoanalytic-psychological propositions without systematically differentiating among them (Rappaport, 1959, p. 16).

One writer, while giving recognition to the importance of the concept of identity for understanding the problems of maintaining personal and social integration in a rapidly changing society and culture, hastens to warn that, "Since at least seventeen various meanings of identity may be distinguished, it is important that its usage in a particular context be clearly determined, lest it become a mere 'fad' word (Becker, 1957, p. 828).

The present study attempts the difficult task of translating this abstract, global, imprecisely defined concept into concrete operational terms. Erikson's clinical descriptions of identity diffusion and his discussion of the healthy personality were closely examined.

In one of his earliest writings on the topic, Erikson describes the central loss of a feeling of identity among young war veterans:

> What had broken down . . . was a sense of identity, a sense of who one is, of knowing where one belongs, of knowing what one wants to do. I must emphasize from the outset that this kind of identity is not the same as that which is called 'personal identity.' It is not an amnesia of one's name or history. It is rather a breakdown of the sense that there is continuity and sameness and meaning to one's life history (Erikson, 1950b, p. 16).

In a later essay Erikson (1959) defines the sense of ego identity as follows:

> The sense of ego identity, then, is the accrued confidence that one's ability to maintain inner sameness and continuity is matched by the sameness and continuity of one's meaning for others (Erikson, 1959, p. 89).

He goes on to describe its concomitants as "a sense of 'knowing where one is going,' and an inner assuredness of anticipated recognition from those who count" (Erikson, 1959, p. 118).

Four aspects of the sense of ego identity were derived from these statements. In contrast with individuals with a diffuse sense of identity: (1) knows who he is, (2) knows where he is going, (3) perceives himself as having "inner sameness and continuity," (4) is certain about the way his perception of himself compares to the perceptions which others have of him.

First, the individual with a well developed sense of identity knows who he is. He feels that he is a distinct person, who differs in his individuality from his family and peer group. On the other hand, the person with a diffuse sense of ego identity tends to feel that he is not fully a person. He has difficulty in sorting out the variety of roles he does or could possibly fill in the group, and is likely to display great variability in his attitude as to just what person he is or wants to be. The expression: "I'm not myself." is typical of this attitude that we all display to some degree.

A second characteristic of the person who has a well developed sense of ego identity is that he knows where he is going. He feels that his past has meaning in terms of his future. At the end of childhood with the formation of an integrated sense of identity a person's selfhood takes on a new dimension that was not characteristic of the earlier stages of development. The addition of long range purposes and distant goals give new meaning to his identity sense. Allport (1961) observes: "Until youth begins to plan, the sense of self is not complete" (Allport, 1961, p. 121). The youth whose life lacks any defining objectives does not have a fully developed sense of identity. This is not to say that his goals have to be rigidly focused, but only that he see some central theme of striving in his existence.

A third aspect of this basic attitude is that those with an integrated sense of ego identity perceive themselves as having "inner sameness and continuity." The personality changes, yet there is a thread connecting it episode to episode--a feeling of being the same person one has been. The past life is continuous with the present. The person with a more confused or diffuse sense of identity tends to show less stability in his self attitudes. He does not feel that he is the same person from day to day, and may show great variability in his behavior and sentiments.

The fourth quality of identity consists of the certainty the individual has about the "stimulus value" he has for others. As opposed to the individual with a diffuse identity-sense, those who are strong on this variable feel certain about their predominant personal characteristics and the way their perceptions compare to the perceptions which others have of them. With a strong sense of ego identity, a person feels clear about his position in the group and that his perception of his self is matched by a sameness and continuity of the meaning he has for others.

A finding of significant correlation among measures relating to these four aspects of identity derived from Erikson would give empirical support to the construct of identity. Therefore our first hypothesis is:

> Hypothesis 1: Measures of the four aspects of identity will be significantly intercorrelated.

Correlates of a Sense of Identity

An integrated sense of identity may be viewed as dependent upon a consolidation of gains made at previous stages of development in which the characteristic problem of these stages have been resolved with a high degree of success. This is consistent with Erikson's statement that "the derivatives of earlier relative achievements . . . become part and parcel of the struggle for identity" (Erikson, 1959, p. 141). Thus far we have been principally concerned with the basic attitude characterizing stage V, Adolescence, at which the psychosocial crisis is "Identity vs. Identity Diffusion." Now we shall turn backward and will look at two of the preceding stages and the attendant attitudes developed therein.

A Sense of Identity and a Sense of Trust

According to Erikson's theory, the "task" of the infant at the first stage of life is the development of a sense of trust. The sense of trust--in himself and the environment--develops out of the child's relationships with the mother. He learns to trust or mistrust his mother as a provider of food, care and love. "Mothers create a sense of trust in their children by that kind of administration which in its quality combines sensitive care of the baby's individual needs and a firm sense of trustworthiness within the trusted framework of their community's life style" (Erikson, 1950b, p. 107). The danger of this stage, most acute in the second year of life, is that discontinuities in care may increase a natural sense of loss, as the child gradually recognizes his separateness from his mother, to a basic sense of mistrust that may last through life.

The development of a sense of trust forms the basis in a child for a sense of identity "which will later combine a sense of being 'all right,' of being oneself and of becoming what other people trust one will become" (Erikson, 1950b, p. 107). A person who resolves the trust versus mistrust crisis with success is more likely to be able to negotiate the problem of identity versus identity diffusion, since the integration of a sense of ego identity can be seen as dependent upon the degree of relative progress made at this earlier stage. This leads us to the following hypothesis:

> Hypothesis 2: A sense of identity will be positively related to a sense of trust.

A Sense of Identity and a Sense of Industry

In the next stage Erikson proposes a conflict of "Shame vs. Autonomy," and as the conflict of the third stage of development he contrasts "Guilt vs. Initiative." The basic attitudes of these two stages are not of the same bipolar type as trust versus mistrust or identity versus identity diffusion. In addition, as contrasted with other personality theorists, Erikson treats shame and guilt as two distinguishable--although not opposite--attitudes which at times may be connected, but are not always so related. Since the derivation of these attitudes does not seem as convincing, and since in some successful empirical treatments guilt and shame do seem to be bipolar rather than independent attitudes, we shall skip these two stages and look to the fourth stage in which the sense of industry develops.

This is the stage that corresponds to elementary school age and the time when the child learns "the prerequisites and is given the opportunity and the life task of developing a sense of workmanship and work participation" (Erikson, 1959, p. 128). In learning to accept instruction and to win recognition by producing "things" he opens the way for work enjoyment. The sense of industry then refers to "a sense of being useful, of being able to make things and make them well and even perfectly" (Erikson, 1959, p. 86).

The danger at this stage is the development of a sense of inadequacy and inferiority in a child who does not receive recognition for his efforts. Thus the conflict is "A Sense of Industry vs. Inferiority."

An insufficient solution of this conflict may result in lasting harm for the sense of identity. It is necessary to establish a good relationship to the world of skills and to those who teach and share the skills in order to consolidate one's social roles. As contrasted with young people with an integrated sense of identity Erikson finds that his cases of severe identity diffusion regularly also suffer from an acute upset in the sense of industry. They often display an inability to concentrate on required or suggested tasks. We would predict, then, that:

> Hypothesis 3: A sense of identity will be positively related to a sense of industry.

A Sense of Identity and a Sense of Intimacy

Thus far, we have been examining the relationship of stage V to the preceding stages and the development of the sense of identity as it relates to the proper development of attitudes at these earlier stages. Erikson asserts that each stage is systematically related not only to the preceding stage but also to the subsequent stages. Although each part of the psychosocial personality comes to ascendance and finds its more or less lasting solution at the conclusion of "its" stage, the development of each basic attitude begins long before its particular crisis. We shall now enter the area of psychosocial elements which are not derivatives, but precursors, of future psychosocial crisis.

The core conflict immediately following that of "Identity vs. Identity Diffusion" is "Intimacy vs. Isolation." Identity and intimacy are closely related, for it is only after a reasonable sense of identity has been established that real intimacy with any other person or even with oneself is possible. The youth who is not sure of his identity avoids interpersonal intimacy--he isolates himself. In discussing this, Erikson says:

> True 'engagement' with others is the result and the test of firm self-delineation. Where this is still missing, the young individual, when seeking tentative forms of playful intimacy in friendship and competition, in sex play and love . . . he is apt to experience a peculiar strain, as if such tentative engagement might turn into an interpersonal fusion, amounting to a loss of identity, and requiring, therefore, a tense inner reservation, a caution in commitment. Where a youth does not resolve such strain, he may isolate himself and enter, at best, only stereotyped and formalized interpersonal relations (Erikson, 1959, pp. 124-125).

Some data exist in support of the relationship of intimacy and identity. As was described above, Yufit (1956) found his "Intimate" group rated higher on identity than his "Isolate" group. We shall further examine this relationship in making the following prediction:

Hypothesis 4: A sense of intimacy will be positively related to a sense of identity.

A Sense of Identity and Authoritarianism

In a defense against a sense of identity diffusion, young people may become "remarkably clannish, intolerant, and cruel in their exclusion of those who are 'different,' in skin color or cultural background, in tastes and gifts, and often in entirely petty aspects of dress and gesture selected as *the* signs of an in-grouper or out-grouper" (Erikson, 1959, p. 92). Erikson asserts that these authoritarian propensities are unavoidable and that, "It is difficult to be tolerant if deep down you are not quite sure that you are a man (or a woman), that you will ever grow together again and be attractive, that you will be able to master your drives, that you really know who you are, that you know what you want to be, that you know what you look like to others, and that you will know how to make the right decisions without, once and for all, committing yourself to the wrong friend, sexual partner, leader, or career" (Erikson, 1959, p. 93). In other words, all young people go through a stage where their sense of identity is more diffuse than integrated, and in such a state they are likely to be authoritarian and intolerant. This is consistent with Jane Loevinger's (1959) point that authoritarianism is a stage of development through which we all go. Young people with less well developed senses of ego identity display more authoritarian attitudes in an attempt to defend against the feelings of diffusion than those individuals who have a more integrated sense of identity and have no need for such a defense. We would predict then:

> Hypothesis 5: Authoritarianism will be negatively correlated with a sense of identity.

Method

Construction of Scales

For each of the four aspects of identity defined above, an eight item Likert-type scale was designed. Each of the thirty-two declarative statements was written following the assumption that if a person strongly agrees with such statements, it would indicate one extreme of the particular characteristic being tapped, and if strongly disagrees, that it would indicate that he possesses the opposite extreme. In order to minimize the effects of acquiescent response set, each scale was composed of an equal number of positively worded and negatively worded items. These items grouped according to trait and directionality are listed in Appendix A. Eight items--four positive and four negative--for each of the other basic attitudes of trust vs. mistrust, industry vs. inferiority and intimacy vs. isolation were also constructed. These are listed in Appendix B.

A reversed version of the F scale provided a measure of authoritarianism. This scale was developed by Christie, Havel, and Seidenberg (1958) and is composed of ten positively worded items and ten negatively worded items.

Five levels of response were provided for all of the items: "strongly agree," "agree," "undecided," "disagree," and "strongly disagree." For positively worded items, "strongly agree" was scored 5, "agree," 4, "undecided," 3, "disagree," 2, and "strongly disagree," 1. Scoring of negative stems was in the reverse direction. In the administration of the scales, items from each of the scales were randomly distributed in the extended opinion inventory comprised of the seventy-six items making up the eight scales. Scores on each variable were arrived at by adding up the reversed scores for the particular items composing each of the eight scales.

Sentence-Completion Instruments

Eight sentence stems for each of the four aspects of identity (total of thirty-two) were written. The stems were specifically designed to elicit responses relevant to expressions of one of the four characteristics of identity (See Appendix C). In addition eight stems were constructed for each of the other three Erikson variables. The stems for trust vs. mistrust, industry vs. inferiority, and intimacy vs. isolation are presented in Appendix D.

The fifty-six stems were assembled in an alternation format and labeled "Incomplete Sentence Blanks." This instrument was administered as part of the same questionnaire containing the scales described above. The stems were preceded by these instructions:

> Complete these sentences to express your real feelings. Try to do every one. Be sure to write a complete sentence.

A scoring-by-example manual was developed for scoring the "Incomplete Sentence Blank" following much the same procedures detailed by Renner, Mahr, and Campbell (1962). On the basis of preliminary administrations of the instrument, 185 protocols were obtained and were used as a reservoir from which responses were abstracted which seem relevant to the descriptions of each of the seven variables. Each response was assigned a weight of

5, 4, 3, 2, or 1 depending upon whether E judged that (5) it strongly indicated the positive pole of the variable, (4) it was somewhat indicative of the positive pole of the variable, (3) it was ambiguous or could not be scored, (2) it is somewhat indicative of the negative pole of the variable, (1) it strongly indicated the negative pole of the variable. Each completion was judged only in terms of characteristics of the variable its stem had been designed to elicit.

This pool of items was assembled into a scoring-by-example manual. The end product was a manual made up of illustrations of scorings of sentence completions for each incomplete sentence stem for the one variable the stem had been designed to tap. (See Appendix E for copy of manual.)

Two raters independently scored samples of 20 protocols for each of the seven variables, the samples being drawn from the pool of 705 questionnaires used in this study. To reduce the "halo" effect as much as possible, each item was scored on all 20 questionnaires before the rater proceeded to score the next item. The stack of questionnaires was shuffled between the scoring of each item. After this procedure had been completed inter-rater reliability was computed for each of the variables by computing the product-moment correlation for raw scores. The correlations obtained were: Knows who he is, +0.95; Knows where he is going, +0.96; "Inner sameness and continuity," +0.96; "Knows his stimulus value," +0.96; Trust vs. Mistrust, +0.98; Industry vs. Inferiority, +0.95; and Intimacy vs. Isolation, +0.97.

In the major validity study, the description of which follows, each of the 705 protocols was scored by one rater, with similar precautions against a "halo" effect. The scores assigned to each respondent on the sentence completion were the sums of the rater's scores on those completions for the stems of each variable. These seven sums were used in the main analysis.

Subjects

The instruments previously described were administered as part of a longer questionnaire (See Appendix F for specimen copy of the entire questionnaire) to 715 male freshmen entering Lehigh University in the Fall of 1963 during Orientation Week. Seven-hundred-and-five questionnaires were returned sufficiently complete to be used in this study and the analysis of data presented here is based upon this sample of 705 male college students.

Data-Processing Procedures

Each of the eight sets of scale items was coded and punched into IBM cards and verified. The scorings for the 56 sentence completions were likewise punched into IBM cards and verified.

In order to facilitate the speed and accuracy of punching the scales for the 705 questionnaires, all items were originally scored in the same direction (A = 1, a = 2, ? = 3, d = 4, D = 5). A program[1] was written for the IBM 709 computer which properly reversed the codings of the positive items (A = 5, a = 4, ? = 3, d = 2, a = 1) and punched them out in a corrected raw data deck. In addition this same program punched out a new data deck composed of a summary card for each of the 705 subjects containing the subtotals of the positive and negative stems and the total score on each of the eight variables. This same subroutine was used in producing a deck of 705 summary cards containing the total scores for each respondent on the sentence-completion measures. (Sentence-completion codings required no reversing of scores.)

---

[1] This program and other subroutines were written by Richard C. Hannenberg.

The corrected data deck containing the scale item scores and the sentence-completion raw data deck were used in separate reliability analysis making use of a modification of a program whose writing by staff at the Northwestern University Computing Center had been commissioned by Professor D. T. Campbell. In order to feed all the data on the two decks of cards into the IBM 709 most expediently, they first were transferred to tape, and then the total scores and positive and negative subtotal scores were intercorrelated producing a set of large triangular correlation matrices. A modification of a Northwestern Statistical Program was employed in these computations.

Results and Discussion

Reliability

Reliability was estimated by computation of Kuder-Richardson ("Formula 20")reliability coefficients. The reliabilities for all the variables under consideration as measured by the Likert-type scales and sentence completion methods are presented in Table 1.

These reliabilities appear to be quite low. In order to check whether or not they are significantly different from zero, the obtained reliabilities could be turned into an F ratio by the formula: $r_{kr} = 1 - \frac{1}{F}$. This F is for the main effect of persons tested against the person-times-items interaction. The degrees of freedom are as follows: (Person -1) and(Persons -1) (Items -1). Since we have 705 persons and 8 items, the degrees of freedom would be 704 and 4,928. From an F table, we discover that the F required for $p < .01$ with these df is approximately 1.13, which translates into a reliability of .11 as the minimum $p < .01$ level. The more persons, and the more items the lower reliability required for significance. Applying this criterion the obtained Kuder-Richardson reliabilities are all significantly different from zero.

However, since the magnitudes of these reliability coefficients compare unfavorably with those usually reported, one possible explanation should be mentioned. An explanation lies in the nature of the concepts themselves, the definition of which poses complex multidimensional constructs. Since increasing complexity of a trait increases the probability that individuals will exhibit unique

Table 1

Kuder-Richardson Reliability Coefficients for Scales and Sentence-Completion Measures

	Scales	Sentence-Completion Measures
Knows who he is	.32	.28
Knows where he is going	.39	.55
"Inner Sameness and Continuity"	.48	.31
"Knows his Stimulus Value"	.23	.33
Trust vs. Mistrust	.46	.43
Industry vs. Inferiority	.60	.29
Intimacy	.43	.46
F scale	.39	

patternings of the component elements, it is generally true that the more complex the measure, the lower will the reliability estimate be. It is important to note that the Kuder-Richardson reliability coefficient for the F scale is only .39 and using that classic instrument as a standard of comparison these new measures compare very favorably.

It can be further argued that internal consistency is irrelevant to the testing of the hypotheses. These items represent an effort at translating some global, abstract concepts into concrete operational terms. The combination of the items is inevitably better at representing the construct in question than any one of them would be. Although in each item some of the specific details introduce tangential values, these irrelevancies differ from item to item, tending to be overweighed, in the total score, by the common core. The item set, whether or not it were to turn out that the individual differences within a measure cohere as a psychological syndrome, would remain, subjectively, the most accurate operational representation of the construct in question.

Relations Between Aspects of Identity

Table 2 presents the correlation matrix which resulted from intercorrelating the total scores of the four aspects of identity, trust, industry and intimacy as measured by the specially constructed Likert-type scales described above with the total scores of sentence-completion measures of these same variables. Such a matrix of intercorrelations resulting when each of several traits

Table 2

Multitrait-Multimethod Matrix Based on Direct-Attitude and Sentence-Completion

Measures of Identity Variables and Predicted Correlates[a]

		Direct-Attitude Method							Sentence-Completion Method						
		A_1	B_1	C_1	D_1	E_1	F_1	G_1	A_2	B_2	C_2	D_2	E_2	F_2	G_2
Direct-Attitude Method															
Knows who he is	A_1	(.32)													
Knows where he is going	B_1	.39	(.39)												
Inner sameness & continuity	C_1	.01	.02	(.48)											
Social stimulus value	D_1	.43	.21	-.06	(.23)										
Trust vs. Mistrust	E_1	.39	.35	-.03	.26	(.46)									
Industry vs. Inferiority	F_1	.36	.37	-.27	.33	.46	(.60)								
Intimacy vs. Isolation	G_1	.22	.24	-.24	.25	.36	.44	(.43)							
Sentence-Completion Method															
Knows who he is	A_2	.23	.19	.10	.24	.11	.08	.03	(.28)						
Knows where he is going	B_2	.15	.25	.01	.10	.13	.15	.10	.25	(.55)					
Inner sameness & continuity	C_2	.16	.13	.32	.18	.07	.06	-.03	.10	.02	(.31)				
Social stimulus value	D_2	.18	.16	.05	.24	.03	-.01	-.01	.24	.15	.10	(.33)			
Trust vs. Mistrust	E_2	.23	.27	.10	.21	.44	.19	.14	.18	.18	.07	.10	(.43)		
Industry vs. Inferiority	F_2	.22	.20	.05	.15	.10	.17	.12	.24	.23	.08	.22	.20	(.29)	
Intimacy vs. Isolation	G_2	.06	.21	.01	.05	.13	.08	.28	.16	.29	.02	.21	.18	.19	(.46)

[a] For df = 703, r = .07 is significant at 5% level; r = .10 at 1% level

or constructs is measured by each of several methods has been referred to as a multitrait-multimethod matrix (Campbell & Fiske, 1959).

This matrix is relevant to examining the prediction of Hypothesis 1 that the four aspects of identity which were derived from Erikson's theory are significantly interrelated.

Campbell and Fiske have noted that: "Insofar as the traits are expected to correlate with each other, the monomethod correlations will be substantial and heteromethod correlations between traits will also be positive" (Campbell & Fiske, 1959, p. 104). Reversing their emphasis on discrimination between measures of different traits, we are concerned here with showing the convergence of related traits (or aspects of a construct) across methods. If traits are predicted to be closely interrelated, they should be significantly intercorrelated when measured by the same methods and when measured by different methods.

Three of these four aspects of identity as measured by the same method and as measured by different methods show significant intercorrelations in all cases. In the single method triangle representing the intercorrelations of the multiple variables measured by the one method of Likert-type scales, "Inner sameness and continuity" is not significantly correlated to "knows who he is," "knows where he is going," or "social stimulus value." Its correlation with the first two of these is near zero and the correlation with the last is negative and almost at 5% level of significance. In the sentence-completion triangle it is significantly correlated with "knows who he is" and "social stimulus value" but again is near zero in its correlation with "knows where he is going."

While all four of the defined characteristics of identity have monotrait-heteromethod values which are statistically significant from zero and hence evidence convergent validity, only "inner sameness and continuity" has a validity diagonal value higher than values lying in its column and row in the multiple-methods block. This would indicate that "inner sameness and continuity" is more closely related to itself measured by different methods than to the other aspects of identity.

Consistent with the preceding evidence of independence "inner sameness and continuity" alone meets the Campbell and Fiske discriminant validity requirement of correlating higher with itself than other traits which happen to employ the same method. "Inner sameness and continuity" contrary to the theoretical prediction would appear to be a trait unrelated to the other three aspects of identity.

The data supports this prediction that the college student respondents studied here who know who they are, know where they are going and are aware of how their perception of themselves compares to the perceptions others have of them, but apparently don't necessarily see themselves as having "inner sameness and continuity" as it has been empirically defined here. Conversely, respondents who are not sure who they are, are also not sure where they are going, or what their "social stimulus value is;" but they don't necessarily lack a feeling of inner sameness and continuity.

Validity of Trust, Industry and Intimacy Measures

Unlike the four aspects of identity just examined, the other Erikson variables of Trust, Industry and Intimacy were not viewed as different aspects of the same basic attitude, but represent distinct, although related, constructs. Discriminant validity is of more concern here since these are presumably different traits which should be distinguishable from one another.

All three traits demonstrate convergent validity since the heteromethod-monotrait correlations for all three are significantly greater than zero. Trust and Intimacy, but not Industry, satisfy the heteromethod discriminant criterion, since these respective validity correlations exceed the correlation of that measure with scores for different traits obtained by different measurement methods (heteromethod-heterotrait correlations).

With one exception, trust satisfies the more stringent monomethod discriminant criterion which requires that the convergent correlations of a measure exceed the correlations of the measure with scores of different traits which share the same measurement method (monomethod-heterotrait correlations). Intimacy tends to meet this stringent discriminant criterion. Industry clearly fails to do so. Such failure may indicate that unwanted "methods factors" have resulted in excessively high interrelationships of scores obtained by a common procedure of measurement.

Thus, trust and intimacy meet the Campbell-Fiske criteria fairly satisfactorily, while industry shows convergent validity but fails to demonstrate discriminant validity.

Identity and Trust

Hypothesis 2 predicts a positive relationship of identity to trust. Examination of Table 2 indicates that the three interrelated aspects of identity are positively related to trust when measured by the same type of method and also when measured by different methods. All of these correlations are significant except for one, the heteromethod correlation between "social stimulus value" and trust which is positive but not significantly different from zero.

It would seem on the basis of these correlations that the individual who resolves the trust versus mistrust crisis on the positive side is likely to be successful in negotiating the crisis at the identity stage. Conversely, those who don't make much progress in learning to trust themselves and the world are less likely to know who they are, where they are going and perhaps to be sure how their view of themselves compares to those around them.

Identity and Industry

Our third hypothesis predicted that identity and industry would be positively related. Because of the failure to demonstrate discriminant validity for a sense of industry some doubt exists as to whether industry can be distinguished from other traits not intended to be measured by its operational specification. It correlates significantly positive with the three interrelated identity characteristics in the monomethod triangles, and is significantly negatively correlated with the Likert-type scale measure of "inner sameness and continuity." In the hetero-method block in Table 2 it can be seen that it is unrelated to "social stimulus value."

The results here remain ambiguous, since the measures of a sense of industry are in question, and the intercorrelation with the defining characteristics of identity yields mixed results.

Identity and Intimacy

The fourth hypothesis to be tested dealt with the relation of a sense of intimacy and a sense of identity. It was predicted that intimacy would be positively related to identity. An examination of Table 2 indicates a significant relationship when the two traits are measured by the same method if one omits the maverick trait of "inner sameness and continuity." In the scale method triangle we find that the latter is significantly negatively correlated with the scale measure of intimacy versus isolation. However, this significant negative correlation is not repeated with the sentence-completion measure, nor between different types of measures. "Social stimulus value," while displaying a comparatively high positive correlation with intimacy in the same methods correlations, correlates below the level of significance in the multiple methods relations.

Our results with respect to Hypothesis 4, therefore, are also mixed, and it cannot be clearly determined whether identity as it is defined here is significantly related to intimacy.

Identity and Authoritarianism

Our last hypothesis predicts a negative correlation between the Christie, Havel, and Seidenberg (1958) measure of authoritarianism and our measures of identity. The prediction was that late adolescents low on identity would be high on authoritarianism and conversely that those individuals who were high on identity would be low on authoritarianism.

An examination of Table 3 shows that the correlations between the F scale and the identity scales are not as predicted. Three are positive and the one negative correlation is not significant. Inner sameness and continuity which failed to correlate significantly with the other measures of identity when both were measured by scales interestingly correlates significantly with the F scale at beyond the .001 level. An examination of the scale items indicates some similarity in items. As defined in the Likert-type scale, inner sameness and continuity may be strongly loaded with a rigidity or inflexibility theme which would explain some of the problems with this translation of the Erikson concept, and its positive correlation with authoritarianism.

The correlations between the F scale and the sentence-completion measures are all in the right direction but fail to reach significance. Our hypothesis of a negative relationship between identity and authoritarianism is not supported.

Table 3

Correlations Between Authoritarianism and Aspects of Identity

	Authoritarianism
Likert-Type Scale	
Knows who he is	**-.07**
Knows where he is going	**+.03**
Inner sameness and continuity	**+.18**
Knows his stimulus value	**+.08**
Sentence-Completion Method	
Knows who he is	**-.05**
Knows where he is going	**-.05**
Inner sameness and continuity	**-.06**
Knows his stimulus value	**-.06**

For *df* = 703, r = .08 is significant at 5% level; r = .10 is significant at 1% level

Summary and Conclusions

The main purpose of this study was to translate Erik Erikson's concept of identity into useful operational terms. Four aspects of identity versus identity diffusion were derived from Erikson's discussion of his theory of personality. In contrast with individuals with a diffuse sense of identity, the person who has achieved an integrated sense of identity: (1) knows who he is, (2) knows where he is going,(3) perceives himself as having "inner sameness and continuity," (4) is certain about the way his perception of himself compares to the perceptions which others have of him. Two types of measures were constructed for each of these elements of identity as well as for three other Erikson variables: trust vs. mistrust, industry vs. inferiority, and intimacy vs. isolation. The two types of measures were a set of Likert-type scales and some specially designed incomplete sentence blanks. These measurement techniques were applied to college freshmen, an age group which should have showed the greatest variation in degree of identity integration versus diffusion.

It was predicted that the measures of the four aspects would be intercorrelated. The data supported the prediction for three of the components of identity. "Inner sameness and continuity," however, appeared to be unrelated to the other measures, while the other three aspects of identity were significantly interrelated.

The meaning of this particular empirical result like the result of any validity study is ambiguous. Any such study is both a study of the instrument and the theory that predicts the results, and for this reason both discriminant as well as convergent evidence were presented. The principal analysis consisted of a multitrait-multimethod matrix following the Campbell-Fiske technique for examining construct validity.

Results lend support to the concept of identity as a significant variable, descriptive of variations in self-attitude among late adolescents and related to the earlier development of a sense of trust versus mistrust. Authoritarianism appeared unrelated to identity, as did a sense of industry versus inferiority. The relationship of intimacy versus isolation to identity was left ambiguous by this study, although a previous study had found the two traits related.

References

Allport, G. W. *Pattern and growth in personality*. New York: Holt, Rinehart & Winston, 1961.

Becker, A. M. Childhood, society and identity: On Erik Erikson's childhood and society. *Psyche. Heidel*., 1957, 11, 821-830.

Butler, J. M., & Haigh, G.V. Changes in the relation between self-concepts and ideal concepts consequent upon client-centered counseling. In C. R. Rogers, & Rosalind F. Dymond (Eds.), *Psychotherapy and personality change*. Chicago: Univer. Chicago Press, 1954. Pp. 55-75.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull*., 1959, 56, 81-105.

Christie, R., Havel, Joan, & Seidenberg, B. Is the F scale irreversible? *J. abnorm. soc. Psychol*., 1958, 56, 143-159.

Erikson, E. H. *Childhood and society*. New York: Norton, 1950. (a)

Erikson, E. H. Growth and crisis of the "healthy personality." In M.J.E. Senn (Ed.), *Symposium on the healthy personality, Supplement II: Problems of infancy and childhood*. New York: Josiah Macy, Jr. Foundation, 1950. (b)

Erikson, E. H. *Young man Luther: A study in psychoanalysis and history*. New York: Norton, 1958.

Erikson, E. H. Identity and the life cycle: selected papers. *Psychol. Issues*, 1959, 1, No. 1.

French, J.R.P., & Kahn, R.L. A programmatic approach to studying the industrial environment and mental health. *J. soc. Issues*, 1962, 18, 1-47.

Gruen, W. Rejection of false information about oneself as an indication of ego identity. *J. consult. Psychol*., 1960, 3, 231-233.

Guilford, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.

Holt, R. R. Recent developments in psychoanalytic ego psychology and their implications for diagnostic testing. *J. proj. Tech.*, 1960, 24, 254-266.

Loevinger, Jane. A theory of test response. *Proceedings of the invitational conference on testing problems.* Princeton, N.J.: Educational Testing Service, 1959.

Monroe, Ruth L., *Schools of psychoanalytic thought.* New York: Dryden, 1955.

Rappaport, D. A historical survey of psychoanalytic ego psychology. *Psychol. Issues*, 1959, 1, 5-17.

Renner, K. E., Maher, B.A., & Campbell, D. T. The validity of a method for scoring sentence-completion responses for anxiety, dependency and hostility. *J. appl. Psychol.*, 1962, 4, 285-290.

Ross, A. O. Ego identity and the social order: A psychosocial analysis of six Indonesians. *Psychol. Monogr.*, 1962, 76, (Whole No. 23).

Rotter, J. B., & Rafferty, J. E. *The Rotter incomplete sentence blank manual.* New York: Psychological Corp., 1950.

Yufit, R. I. Intimacy and isolation: Some behavioral and dynamic correlates. Unpublished doctoral dissertation, Univer. of Chicago, 1956.

Appendix A

Items for Scale of a Sense of Ego Identity

Defining Characteristics:

A. Knows who he is

Positive Items

(4)* It isn't necessary to be a chameleon and be all things to all people in order to get ahead in life.

(20) Although there is a lot of talk nowadays about the loss of individuality in our mass society, most of us feel that we are distinct individuals in our own right.

(25) Even if it were possible to start life all over again, most of us would decide to be the same persons we are.

(69) Most people know who they are and where they belong.

Negative Items

(10) The worst thing that can happen to a person is to turn out to be a nobody.

(43) At times, nearly everyone would like to be someone else.

(64) What a bore it is, waking up in the morning always the same person.

(75) It is only human to worry a lot about finding yourself.

B. Knows where he is going

Positive Items

(8) The major decisions a person makes are guided by the plans he has for the future.

(40) Most people can plan their lives so that they get what they want.

(52) Most of us have a fairly good idea what we'll be ten years from now.

(63) The present has meaning in terms of the future.

Negative Items

(1) Life is chaotic, without direction or meaning.

(23) The most difficult task for a person is to decide what kind of life they want.

*Number in parenthesis indicates item number in the extended opinion inventory.

(65) Most people don't know where they are headed.

(72) It is the exceptional person who plans for the future.

C. Perceives himself as having "inner sameness and continuity"

Positive Items

(15) A person doesn't change much once he has started out in the world.

(28) A man's present life is continuous with his past.

(32) Life is guided by certain principles one learns as a child.

(59) A man's tastes and sentiments remain relatively constant.

Negative Items

(48) Most people act and express themselves quite differently at different times.

(56) A person's own estimates of his capabilities are likely to change quite frequently.

(61) No one is the same person from day to day.

(70) Every man has felt that he would like to cut his moorings and quit his home, his parents, and his friends.

D. Is certain about the way his perception of himself compares to the perceptions which others have of him

Positive Items

(4) It isn't necessary to be always proving yourself to others, it is enough to know it yourself.

(18) One doesn't depend on others to tell him who he is, and where he is going.

(29) A person can be confident of getting recognition from those who count.

(50) A man can be sure of himself regardless of the opinions of others.

Negative Items

(22) An individual should adjust his thinking about himself to concur with the views of the majority.

(33) What really matters is what other people think; it is not enough just to be sure of oneself.

(36) Most of us are not pleased with the "type" of person we appear to be to others.

(74) Even when they are quite alone, most people worry a good deal about what others think of them.

Appendix B

Items for Scale of a Sense of Trust

Positive Items

(35) A man can place his trust in the future.

(38) It is foolish to worry about your own capabilities, because most of us can be trusted to come through when it really counts.

(66) When you come right down to it, most people will help you when you are in trouble.

(68) The older you get, the more you realize that you have to trust other people.

Negative Items

(12) If given a chance, most people would take advantage of you.

(41) In times like these, no one can trust what the future will bring.

(47) The only thing to trust is mistrust.

(58) It is rarely safe to trust a stranger.

Items for Scale of a Sense of Industry

Positive Items

(2) There is great pleasure in giving a job all you've got.

(55) Distant goals deserve more effort than momentary pleasures.

(62) A person who can do a job doesn't have to feel inferior about anything.

(67) A man should work hard when he works.

Negative Items

(5) Many people don't have what it takes to finish a job.

(16) People wouldn't do any work at all if they didn't have to.

(26) If you are not getting results on a job, it is best to quit before you expend too much energy.

(45) Most young people today feel it is foolish to do more work than is necessary to "get by."

Items for Scale of a Sense of Intimacy

Positive Items

(6) We arrive at wisdom through our intimacies with other people.

(14) A close physical relationship in marriage plays an important part in one's happiness.

(21) We all need the love of those who are close to us.

(49) It is better to have a few close friends than many acquaintances.

Negative Items

(7) Familiarity breeds contempt.

(17) It is best not to become very intimate with another person, because it always leads to frustration and disappointment.

(19) The man who walks in solitude is the one who experiences the greatest enjoyments of which our nature is capable.

(46) Being too involved with someone else takes something away from yourself.

Appendix C

Sentence-Completion Stems for a Sense of Identity

Defining Characteristics:

A. Knows who he is

(1)* If I could change my name, I . . .

(8) For me to really be a person . . .

(18) When somebody confuses me with someone else, I . . .

(22) I am not myself, when . . .

(29) I feel like an imposter, when . . .

(36) Knowing who I am is . . .

(43) Pretending to be somebody you aren't is . . .

(50) Knowing where a person belongs is . . .

B. Knows where he is going

(3) My goals are . . .

(10) The things that I want out of life are . . .

(17) I am looking for . . .

(24) When I am 30, I expect to be . . .

(31) I feel committed to . . .

(38) My life is built around . . .

(45) In making plans for the future, I . . .

(52) I anticipate . . .

*Number in parenthesis indicates item number in the sentence-completion battery.

C. Perceives himself as having "inner sameness and continuity"

(5) When I compare the kind of person I am now with the kind of person I use to be, I . . .

(12) When new fashions and ideas come along, I . . .

(19) The person I was yesterday and the person I am today are . . .

(26) The things that once were important to me . . .

(33) What I think changes most about a person . . .

(40) From past experience I have learned that I . . .

(47) If it were possible to go back in time and see myself as I used to be, I would probably feel that I . . .

(54) When I try to make sense out of my past life. . .

D. Is certain about the way his perception of himself compares to the perceptions others have of him

(7) I am sure that people think of me as . . .

(14) The opinions of others are . . .

(21) When I think about the kind of person I appear to be to others, I . . .

(28) I feel that most people who meet me for the first time think I am . . .

(35) Proving myself to others is . . .

(42) When I walk into a room and know that everybody is looking at me, I . . .

(49) When someone tells me something about myself, I . . .

(56) In a comparison of the way I see myself and the way others see me, I think . . .

Appendix D

Sentence-Completion Stems for a Sense of Trust

(2) If you trust most people, you are apt to get . . .

(9) A person who has lost his wallet, if he has his name and address in it, can probably expect . . .

(16) When you are having difficulties, you can expect other people to . . .

(23) Without law enforcement, most people would . . .

(30) If you don't know much about something, you can expect other people to . . .

(37) I think the person who trusts in his own abilities . . .

(44) The man who invests his efforts in the future . . .

(51) The best policy when someone offers you something for nothing is . . .

Sentence-Completion Stems for a Sense of Industry

(4) When I find it difficult to carry out a job, I . . .

(11) Whenever there is more work to be done than I anticipated, I . . .

(18) Sometimes I get so preoccupied with some activity . . .

(25) When I have a required task, I . . .

(32) Knowing how to do a job well . . .

(39) When I stop getting immediate results from what I am working on . .

(46) I feel most inferior with regard to my abilities, when I . . .

(53) I think that the anticipation of achievement . . .

Sentence-Completion Stems for a Sense of Intimacy

(6) Intimate relationships for me are . . .

(13) In the presence of women I feel . . .

(20) At small parties, I feel . . .

(27) Working very closely with another person . . .

(34) To be aloof and inaccessible . . .

(41) When I have to be by myself for any length of time . . .

(48) When I am separated from a person I love . . .

(55) I feel that friendship is . . .

Appendix E

Manual for Scoring
Identity, Trust, Industry, and Intimacy
Incomplete-Sentence Blanks

Frank Baker

Four components of the sense of identity versus identity diffusion have been derived from Erikson's discussion of the concept in Childhood and Society, The Healthy Personality, and The Problem of Ego Identity. In contrast with individuals with a diffuse sense of identity, the person who has achieved a sense of identity: (1) knows who he is, (2) knows where he is going, (3) perceives himself as having "inner sameness and continuity," and (4) is certain about the way his perception of himself compares to the perception which others have of him. Eight sentence-completion stems have been prepared for each of these four aspects of identity (total of thirty-two). The stems were specifically designed to elicit responses relevant to expressions of one of the four aspects of identity. In addition eight stems have been constructed for each of three other Erikson bipolar variables; trust versus mistrust, industry versus inferiority, and intimacy versus isolation.

This manual is made up of illustrations of scorings of completions for each incomplete sentence stem on the variable the stem is designed to tap. Each completion is assigned a score of 5, 4, 3, 2, and 1 depending upon whether or not it is judged that: (5) it strongly indicates the positive pole of the variable (4) it somewhat indicates the positive pole of the variable (3) it is ambiguous or cannot be scored on the variables (2) it is somewhat indicative of the negative pole of the variable (1) it strongly indicates the negative pole of the variable.

Knows Who He Is

1. If I could change my name . . .

5 I wouldn't
I wouldn't bother; my name only stands for what I make myself
I still wouldn't, I respect it
I would not because a rose in any other language . . .
It wouldn't change me

4 I don't think I would
I probably wouldn't bother

3

2 I would change my middle name
I would only change my middle name
I would change the Jr. to a II
I would switch the first and middle names

1 I would change it to (some proper name)
I would probably pick a prettier one
I would make it sound more official
I would change it to a more common and better sounding one
I would make it less Jewish
I would make it more American

8. For me to really be a person . . .

5 I'll stay the way I am
I must express the way I feel, no matter what
I would act as I always have
I must achieve all my goals and be ever mindful of my duty to God
is to be myself as I am
I try to be my natural self
comes naturally
is quite normal

4 I have to be an individual
is really not much of a problem for me
isn't too hard

3

2 I must respect and have confidence in myself
I should keep my mouth shut more
I must have friends
I must work at it
I must have my conscience clean and be doing something useful
I need a car and money

1 is my fondest wish
I will have to develop my character
I will have to change my personality
is one of my greatest concerns
I would need to find someone to model myself after
I need other people to help me

15. When somebody confuses me with someone else, I . . .

5 am not bothered by it
don't care
don't worry about it
take it in stride
try to correct them with as little embarrassment as possible
pass it off and try to make it less embarrassing for the person
tell them who I am

4 don't get excited, I do the same thing too
it doesn't do anything to me since I have trouble remembering faces
think it is funny
really don't care

3

2 feel it is a natural mistake
tease them
am usually embarrassed
make a joke of it especially if the someone else is handsome or something

1 let it go if I don't know the person well or thank him for the compliment and correct him
try to decide if it was a compliment in the eyes of the person
ignore the person
wonder who the other guy was; it makes me uneasy
I feel confused
find out who the other guy was
begin to wonder who I really am

22. I am not myself when . . .

5 I am physically ill
I don't feel "up to par"
I am always myself
I don't act nice to others as I usually do
when I am inconsiderate or forgetful
I am on vacation
never

4 I lose my temper
I drink a little too much

3

2 I am with a large group of people
I try to give girls "a line"
I occasionally catch myself acting a little "phony"

1 I try to act like someone I am not
I am emotionally disorganized
I am all alone
I "act the role"
I try to be what other people want me to be

29. I feel like an imposter when . . .

5 I rarely feel like an imposter
I never feel like an imposter
I don't believe in what I have to do
other people try to force me to do things that just aren't me

4 I occasionally act without thinking ahead
hardly ever
I haven't come upon this situation

3

2 I speak of athletic skills
I try to act better than I feel I am
I am with other people
I give a girl the line

1 people refer to me as a "face man" and "a cool stud"
a person thinks someone else is me
I'm not myself
I act like someone I'm not
I play the role
I put on a false identity to impress people
I am one

36. Knowing who I am is . . .

5 helpful in relations to others
easy
not a problem for me
something I have learned
a question I can answer to my satisfaction

4 fairly clear to me
not much of a problem for me
no longer much of a concern

3

2 another goal of mine
going to be important in a few years
sometimes difficult
a problem

1 the hardest thing in the world
impossible
something I don't really know
extremely important for me to find out
the most important question in my life
I desperately need to know

43. Pretending to be someone you aren't is . . .

5 bad because it shows that you feel you are inferior
a lie and will cause trouble later on
a serious mistake
something I don't do
something I feel no need to do

4 only to fool yourself
risky and not right if you mean it for life
very hard to do since you don't fool yourself
unnecessary and foolish

3

2 often a very saddening experience
a mistake I made
stupid, but useful sometimes
fun for a short time
natural, once in a while

1 what I have to do
something I often do
the only way to get along
bad but I have to do it all the time
often very necessary
pleasureable, because I dont' like who I really am
human, everybody does it
easier than being yourself
a good way to change
the one thing that makes me happy
often necessary, you must adapt yourself
to your environment at any one instance

50. Knowing where a person belongs is . . .

5 something I know about myself
helps me to be the person I am
not difficult
easy
no problem
helps him in making his plans

4 not much of a problem if you persevere
seldom of concern

3

2 often difficult, but not impossible
sometimes a real source of worry

1 impossible
important for me to find out
something I wish I knew about myself
good, I guess, but I can't find my place

Knows Where he is Going

3. My goals are . . .

5 well worked out
to get a degree in Biology
to become a lawyer
fairly well defined
clear and well defined
to finish college, get married, be successful in my career, and to raise my children

4 money, success, happiness
to contribute something to mankind
to have fun
pretty clear

3

2 improvement
to be with my future wife and to face whatever comes
not easily defined
somewhat undefined

1 not at all clear
very vague
unimportant
chaotic
without any real meaning
to find myself
nonexistent

10. The things that I want out of life are . . .

5 to be a successful doctor and a good husband and father
a happy family and a successful career
a good job in the biological field
well planned out

4 love, marriage, success, happiness, fun etc.
family, job, to get to Heaven

3

2 everything
sometimes hard for me to decide on
hard to get

1 not easily gotten on my own
are difficult for me to decide about
are not at all clear to me
to find myself
all beyond my reach
non realistic

17. I am looking for . . .

5 my plans to work out
ways of doing the things I plan to
knowledge
success in my chosen career

4 a good job and a fine girl
wife, job and happiness
an opportunity for advancement
a good education

3

2 satisfaction, contentment, peace of mind
meaningful ideas
security
a secure future
a place in life

1 something sound to base myself on
the interest that will hold my life
more people like myself
an answer to who and what will I be
the easy way
who knows what
a purpose

24. When I am 30, I expect to be . . .

5 a businessman with a degree in law
a teacher
a well paid engineer
a medical practitioner
setting up practice in my home town

4 father, successful, happy
married, job, house, well off
established

3

2 starting out in the world
having fun

1 a lot more mature than I am now
I don't know what I'll be
a nobody, I'm afraid
dead
what my parents want me to be, an engineer

31. I feel committed to . . .

5 pursuing my goals and ideals
my professional plans to be a lawyer
my religion
my plans for the future
my family, my God, and myself
doing the best job I can
living life to my fullest capacity

4 achieving happiness for my family and myself
helping other people
a fairly definite pattern in my life
most of my plans for the future

3

2 my former and only real girl friend
act the way one my age should act
not very much
hardly anything or anybody at all

1 live up to everyone's expectation of me
nothing
what others expect of me
what my parents want me to do
no definite goals

38. My life is built around . . .

5 my plans for my career as an engineer
my thinking and God
my values
my church, my fiancee, my family and school
the principles I learned as a child

4 preparing for the future
abilities and interests
my home
my intelligence

3

2 a rather shaky hope for something else
not too much, I have to admit
some hazy notions about what I want out of life
other people

1 nothing
I don't know what
no particular thing
not much at all
me
fantasies

45. In making plans for the future, I . . .

5 organize things carefully so that the most will be accomplished
want to be an electrical engineer
think of my present needs then future and plan for both
know where I am headed

4 am careful
try to anticipate future conditions
look at my goals and the over-all picture
try to form them around my objectives

3

2 work hard, love strong, and wait for the future
am too optimistic
am concerned about a few indefinite gaps wherein I need education

1 hope to start about now
have many unanswered questions
I am rather indefinite
have done next to nothing
see only death approaching

52. I anticipate . . .

5 my future
graduation, marriage, and life in general
much joy today and in the future
getting what I want out of life
the accomplishement of my goals

4 being fairly successful in what I attempt
not doing badly

3

2 some hardships, but I hope for a happy life
too many things
work for almost everything
a long hard fight to reach my goals

1 nothing
death
that I will fall short of my goals for the future but
not because I didn't try
a chance to put my life in order
the end of the world

Inner Sameness and Continuity

5. When I compare the kind of person I am now with the kind of person I used to be, I . . .

5 see much similarity
am the same
see that basically my personality has not changed
am really the same person
see no difference

4 can see very little real difference
see only slight changes in myself
find a similar but older image
haven't changed much

3

2 am not as virtuous
see progress
see that there is still a long way to go
feel that I have changed
try to change some parts

1 am a much better person in that I know myself better
consider myself an entirely different person
am ashamed of my past life
think I am a fast changing person with different ideas everyday
I am sorry to say I was better

12. When new fashions and ideas come along, I . . .

5 look them over
do what I like
often need proof of their usefulness
stick with the old
stick with what I like

4 take them or leave them
think about them and form an opinion
don't pay much attention
only accept them if they are suited to my taste

3

2 watch for others' reactions
usually try them after a while
usually go along if I like them

1 am quick to adopt them
jump on the band wagon
join in
feel compelled to go along with them
flow with the stream

19. The person I was yesterday and the person I am today are . . .

5 basically the same
one and the same
entirely the same
continuous

4 fairly similar
usually the same person
the same in many ways

3

2 different in some ways and the same in others
generally a little different for the better I hope
only similar

1 not the same
two entirely different persons
never the same
completely different
as different as night and day

26. The things that once were important to me . . .

5 still are
are only slightly different
are relatively the same
haven't changed

4 still have some value depending on what they were
usually are now
are only a little different

3

2 have lost some importance
have less meaning now
have given way to truly important things

1 now seem very unimportant
have changed very greatly
are no longer
now seem childish and trivial
have changed and are changing

33. What I think changes most about a person . . .

5 are minor aspects of his physical appearance
is his body, but his attitudes don't change much
are really not very important
is the color of his hair
is his environment
is his age
a person doesn't really change much, he is basically the same person

4 is what he knows about life
are his accomplishments
is his maturity

3

2 is his consideration toward others
are some of his feelings about himself
are some minor aspects of his personality

1 is what he is
is his personality
are his values and ideals
is his whole view of himself and the world
is the way a person acts
is what he believes
is the kind of person he is
is who he is

40. From past experience, I have learned that I . . .

5 I must try to be myself
a person doesn't change much after all
am still the same person I was and will continue to be
and most other people stay pretty much the same
can build a successful future

4 can do almost anything I want to do badly enough
must be able to stand upon my beliefs to others if necessary

3

2 dont' have enough confidence in myself
make many mistakes

1 act and express myself quite differently at different times
am not the same from day to day
can always be a better person than I am today
can't count on myself
am not much of a person

47. If it were possible for me to go back in time and to see myself as I used to be, I would probably feel that I . . .

5 have changed very little
I was really not very different
would be the same if I could start over again

4 was out of the group but in for the future

3

2 made some stupid mistakes
did some pretty dumb things
could be vastly improved

1 have changed an awful lot
was a better person
have made great progress
was a fink
was a fool and didn't do the things I should have done

54. When I try to make sense out of my past life . . .

5 it is easy
it usually is pretty logical if I consider everything
it all rings true and points to a good life ahead
I can
I see a pattern
I see why I acted as I did
I find it falls in stages
I do

4 I am able to do so within limits
I am reassured
I am fairly successful

3

2 I find it easy except in one area
it is sometimes a puzzle
I see the main areas that in my life were lacking
I can't figure all things out

1 I can't do it
I get utterly confused
I don't do so well
it's impossible
I give up

His Self-Perception vs. Others' Perceptions of Him

7. I am sure that most people think of me as . . .

5 I think of myself
what I am
pretty much the same as I think of myself
they have in the past

4 a person who wouldn't let them down in a pinch
a practical joker who can get serious and is not hasty
an intelligent well-mannered friendly person
an all right guy
nice

3

2 being conceited
an ok guy I hope
a cool guy and a face man
a little off center
a mad genius
unsettled
wild

1 a nobody
unusual
more than I am
just another person without any characteristics
unpredictable
an unusual individual
a confused person

14. The opinions of others are . . .

5 worth my consideration
to be taken just as opinions not facts
not important to me
not really important, if one is sure of oneself
not to be followed as a matter of necessity

4 to be valued and analyzed for their worth
to be taken with a grain of salt
sometimes worth listening to

3

2 important but not essential
meaningful if I respect the person
a pain to me until I can think them over and find if I'm wrong

1 something I worry about all the time
of some good use to me
extremely important to me
a primary consideration
what really matters
music to my ears
as important as my own
my best guide as to what is correct

21. When I think about the kind of person I appear to be to others, I . . .

5 feel satisfied
am not disappointed
am proud and happy
am confident that I will be recognized by those who really matter
am happy cause I am always myself

4 do not get too concerned
think they're wrong but it's all right
am generally satisfied with the image
am not surprised

3

2 find that I have faults and strong points
am ashamed of some things
hope I appear all right
worry
sometimes wish I could change my ways
should change myself a little

1 feel very concerned
feel sick
don't like myself
think I don't live up to their opinions of me
am not the person I seem to be
try to be like they want me to be
really get frightened
wonder if they are right
wonder what they say about me
have a hard time visualizing their image
am very confused

28. I feel that most people who meet me for the first time think I am . . .

5 an individual
a likeable person
intelligent and friendly
as I am
great
a distinct and unusual individual

4 quiet
a quiet reserved person
all right
intelligent
average
am pretty much as I am

3

2 a smooth talker and a face man
much older than I am
better than I am

1 without much personality
a nonentity
a nothing
not worth their attention
just another person
different than I actually am

35. Proving myself to others is . . .

5 not necessary
not something I work at
relatively unimportant to me
unnecessary for I don't need to prove anything

4 really doesn't concern me
not very important to me
not much of a problem for me

3

2 important in some situations
significant but not essential
somewhat difficult
not too important, but it has some importance

1 one of my main occupations
almost a compulsion of mine
a primary consideration
my main goal
a normal thing
a life long task

42. When I walk into a room and know that everyone is looking at me, I . . .

5 ignore it
say Hi
smile
enjoy their attention
am not upset
look right back at them

4 keep walking
feel like saying hello and walking in
straighten my shoulders
am not very concerned

3

2 feel a little self-conscious
feel uncomfortable
wonder why

1 want to sink right into the floor
am embarrassed
feel very nervous
feel miserable
am not myself
feel like walking back out

49. When someone tells me something about myself, I . . .

5 take it for what it is worth
weigh its validity
test it against my own thoughts
am usually not surprised at what they say

4 feel pretty sure I know what they are going to say
listen only if it is logical criticism
analyze it

3

2 sometimes feel hurt
fear he knows more than he says
am likely to feel they don't understand me
usually am interested
try to correct it usually

1 usually agree
accept their opinion and try to change
accept it with much interest
adjust my thinking about myself
am often hurt and surprised
believe him
am very interested
take it seriously
listen!!
try to please him by changing

56. In a comparison of the way I see myself and the way others see me, I think . . .

5 we are in agreement
the two views are nearly the same
the two correspond
I know how the two compare
I portray myself as I really am
we agree

4 that I may overemphasize some parts
there is surprisingly little in the difference that I haven't recognized before
they are close
I know how the two compare most of the time

3

2 I should try to change
I could be a better person
I am more critical than the others
too much of myself
I am sometimes wrong
they rarely see me as I am
they should know me better before they judge
I'm prejudiced

1 they see an entirely different person
I am a flop
I am ashamed of not fulfilling others' ideas of me
I am not the person they think I am
am unhappy about the way others see me
most people don't know what to expect of me
I overestimate myself
I would like to be more like I think I am
there are two of me

Trust vs. Mistrust

2. If you trust most people you are apt to get . . .

5 along fine in the world
along with everybody
a lot out of life

4 treated fairly most of the time
many worthwhile friends
satisfaction
along fine usually

3

2 some trouble, but mostly friends
hurt sometimes
stepped on once in awhile
discouraged

1 a kick in the teeth
taken advantage of
yourself taken
hurt badly some day when someone you trust ins't trustworthy
fooled because not all people can be trusted
swindled, cheated and tricked too often

9. A person who has lost his wallet, if he has his name and address in it, can probably expect . . .

5 to have the wallet returned
the finder to send it back
to have it returned
it back
me to return it

4 to find his wallet
a very good chance of getting it back

3

2 it back without the money in it or not at all
not at all
someone to return it perhaps
a 50% chance of its return

1 to never see it again
not to get it back
to lose it permanently

16. When you are having difficulties, you can expect other people to . . .

5 try and help you
be sympathetic
help you as much as they can
lend a hand

4 usually help you out, depending on the people
give some help
sympathize if they understand the problem
lend a hand up to a point
help if they don't stand to lose by it

3

2 not sympathize with you
listen with false interest
sometimes help, but not usually

1 take advantage of your difficulty
leave you alone
not help you
not be concerned
pass you by
watch
ignore you

23. Without law enforcement, most people would . . .

5 be the same
still conform to the law
set up a law code of their own
behave as they do now
try to preserve order
lead a good life anyway
respect other's rights

4 maintain the basic and more important laws
still remain relatively lawful
go wild for a while then control themselves
trustworthy for the most part

3

2 break many of the minor laws
probably act a little worse in small things
just keep the ones they think are necessary
live in fear
keep themselves in check for a while

1 run amok
become unlawful
do whatever they pleased
hurt other people
run wild like animals
be barbaric

30. If you don't know much about something, you can expect other people to . . .

5 tell you if they know
give you assistance
to be understanding
help you out
help you learn about the subject

4 help you if it doesn't hurt them
help you sometimes
come to your aid much of the time

3

2 give you one side of the story
not to be of any help
kid you about it
listen to you condescendingly
sling the ball about it
give you advice whether they know about the subject or not

1 take advantage of your ignorance
use it against you
ask you about it intentionally
pass you by

37. I think the person who trusts in his own abilities . . .

5 has the right idea
will succeed in life
will go where he wants to go
has the confidence of others
can satisfy his goals
is wise
can succeed in his work

4 is probably right
usually succeeds
succeeds, if he is not overconfident

3

2 has an advantage over someone like me who doesn't
may do a lot of things wrong
can sometimes become lost
may be over confident
could get into trouble

1 is a fool
deceiving himself
will end up in a lot of trouble
shouldn't
is the first person to make a mistake
will realize his mistake in the end

44. The man who invests his efforts in the future . . .

5 can be expected to be well repaid
is rewarded
will get a dividend back
is quite often repaid by his insight
is smart (wise)
is the man who will profit in the long run

4 is usually successful
is successful if he doesn't forget others

3

2 is not always sure of a return
will very likely not be rewarded
plans to live to see the future
sometimes fails

1 is foolish
is likely to get it in the neck
wastes his investment
often doesn't live to see the future
is taking too much for granted

51. The best policy when someone offers you something for nothing is . . .

5 take it
accept graciously
thank them and take it

4 to accept it only if you want it
to evaluate its use and then decide
"don't look a gift horse in the mouth"
take it if you have no reason not to

3

2 to be wary
to study the offer with suspicion
investigate it thoroughly
find out what the trick is

1 not to take it
find out what they are really offering
watch out
not to trust them

Industry vs. Inferiority

4. When I find it difficult to carry out a job, I . . .

5 keep right on trying until I succeed
try harder
try, try again
bear down harder
love it

4 usually refuse to stop
swear, and continue
am irritated, angry

3

2 stall around
do just enough to get by
sometimes get bored
sometimes procrastinate

1 give it up
quit
feel inferior
usually quit
get somebody else to do it for me

11. Whenever there is more work to be done than I anticipated, I . . .

5 still go ahead and do the work
work even faster
generally just work harder; if I start a job, I try to finish it
do it anyway
work longer than I anticipated

4 gripe some and then get to work
am angry, but do it anyway

3

2 usually don't finish it
leave some out
get help
delegate it to others

1 feel cheated and depressed
quit
give up in despair
try and get out of it
got somebody else to do it for me

18. Sometimes I get so preoccupied with some activity . . .

5 that I devote long hours to it
that I work as hard as I can
when I'm trying to do a very good piece of work
because I want to finish what I've started
that I hate to stop, especially if I'm making progress

4 that I spend a pretty good amount of my time on it
but I usually try to keep things in balance
usually a constructive one I hope

3

2 that I get myself in a little trouble
I don't very often
not me, I have trouble concentrating most of the time

1 that I drift a way from the real world
that I lose track of time and forget cares and responsibilities
I lose everything else
I lose my sense of values and go too far
that I become oblivious to the rest of the world
but not if something more interesting comes along

25. When I have a required task, I

5 do it
usually do it right away and get it over with
try and do the best job I can

4 try to set aside enough time to do the job well
am irritated, annoyed, but I do the job
can usually be counted on to complete it in plenty of time for the deadline

3

2 sometimes seek help
am nervous
sometimes find it difficult to get down to work

sometimes find it difficult to get down to work

1 I find it difficult to concentrate
put off doing it
try not to think about it
have difficulty beginning to work
I procrastinate
pretend it is not required

32. Knowing how to do a job well . . .

5 is a good feeling
gives me a wonderful feeling of pride
is worth the effort it took to learn how
is a source of satisfaction

4 is appreciated by others
is something I occasionally get quite a kick out of
is an experience I am trying to enjoy more often

3

2 is a feeling that I rarely have
doesn't always give just rewards
is not all it's cracked up to be

1 is something I wish I knew
is an enviable trait
is my main objective in life
is something that you really don't need to know

39. When I stop getting immediate results from what I am working on . . .

5 I try harder
keep on working at it
I believe this is no reason to stop working
it doesn't phase me

4 I try to analyse what is wrong with my method and then resolve the problem and continue
try a few more times

3

2 I am tempted to give up
become discouraged
get bored
I relax for a while

1 I stop immediately
give up whatever I'm working on
I quit before I waste too much energy
wonder if there is something wrong with me
I am prone to stop working
I wonder why

46. I feel most inferior with regard to my abilities, when I . . .

5 am not doing what I know I can
don't work up to my full capabilities
don't give a job all I've got
don't feel

4 don't usually feel inferior with regard to my abilities
don't do things as well as I usually do
am not as productive as I like to be

3

2 am learning something completely new
take part in athletics
must give an impromptu speech

1 compare myself to others
see people who can do things well
when I can't perform my work
just can't do things right

53. I think that the anticipation of achievement . . .

5 is what keeps me going
is enjoyable
in myself is not unfounded
is the key to success

4 is usually beneficial
is good if the person has the abilities to achieve

3

2 can be detrimental
sometimes blocks the achievement itself
is bad if the person then fails
is all right for some
is counting chickens before they are hatched

1 is a waste of time
is not necessarily worth doing
is to be conceited
is premature and unwarranted
is foolish

Intimacy vs. Isolation

6. Intimate relationships for me are . . .

5 wonderful
necessary
very satisfying to me
very rewarding
something I seek and enjoy
rewarding and I try to seek them out and develop them
with the girl I love

4 usually enduring relationships
with my family only
with a select few
few, but necessary

3

2 fairly short-lived affairs
boring
embarrassing at first
rare
very few
difficult and few

1 overwhelming
frustrating
something I can do without
unnecessary
nonexistent

13. In the presence of women I feel . . .

5 happy
comfortable
wonderful
right at home

4 pretty comfortable usually
fairly happy
usually just fine
swell, most of the time

3

2 a little strained
sometimes uneasy
somewhat upset, depending on the woman
sometimes embarrassed

1 inhibited
shy
ill at ease
nervous
extremely uncomfortable
that I prefer things to be formal
threatened

20. At small parties, I feel . . .

5 right at home
very comfortable
at ease
that I would like to know everyone
like having a good time
like being the life of the party
better than at large parties
close to others, and I enjoy it

4 at home when among friends
fine, if I know a few of the people

3

2 less at ease than at big formal parties
like running away sometimes
sometimes unhappy being with the people I'm with

1 threatened
out of place
all alone
isolated
uneasy, because I don't know how to act
wiped out

27. Working very closely with another person . . .

5 gives me a good feeling
is something I enjoy
is a very satisfying experience

4 gives me pleasure, if I like the person
can be great fun
gives you a chance to build a friendship
gives you a chance to observe another person
is enjoyable at times

3

2 can be very strained
can be dangerous
can be good or bad

1 is something I find very distasteful
leads to conflict
is something I prefer to avoid
gets under my skin

34. To be aloof and inaccessible . . .

5 is wrong
is stupid
deprives a person of one of the richest experiences given to man: friendship
sets one apart from the joys of life
is not for me

4 is something I don't think I'd care for
doesn't sound like very much fun
is rarely necessary or desirable

3

2 is sometimes a good policy
stupid but sometimes necessary
once in a while is very refreshing

1 is what I enjoy
is what I long for
can be one of the greatest pleasures in life
is good for a man

41. When I have to be by myself for any length of time . . .

5 I miss my friends
become depressed at being away from those I love
I feel a need for people
I count the hours
I miss the company of others

4 I feel a little lonely
I think about those who are absent
I am not very happy being alone

3

2 I read and pretty much enjoy myself
I sometimes like it
it doesn't bother me much to be alone

1 I enjoy the feeling of isolation
it is a welcome experience
I am very happy to be by myself

48. When I am separated from a person I love . . .

5 I realize how much I need their love
I can hardly wait to see the person again
I miss them
I try to keep in touch
I am hard to live with

4 I miss them, but I don't go overboard
I miss them, but life goes on
I am not very happy

3

2 I sometimes doubt our love
I sometimes enjoy it

1 I bear it with no difficulty
I don't mind at all
I feel much relieved
I eventually lose that love
I look for another love

55. I feel that friendship is . . .

5 very important
one of the finest experiences people can share
is one of the most important factors in a happy life
good and noble
something I greatly need and enjoy

4 great if they are true friends
great with a select few
usually rewarding

3

2 something I feel strained about
sometimes a bind
a crutch usually

1 not really necessary for happiness
greatly overrated
a bother
unnecessary
impossible for me

Appendix F

. Lehigh University

Self Report Questionnaire

Form I-A

Instructions:

In this questionnaire you are asked to answer some questions about your personal and social attitudes, your background, and your plans. The questionnaire is being given to learn more about students and their experiences in college. It is part of some research which is being conducted in order to help provide the basis for more enlightened planning of college programs and more effective teaching and guidance of students.

Please answer the questions as accurately and frankly as you can. There are no right or wrong answers. Your own personal opinion is what matters. Be sure to answer all parts of the questionnaire, but work rapidly giving your first responses.

Since we have to be able to keep track of who has and who hasn't filled out the questionnaire, we are asking that you write your name at the bottom of this page. A code number will be assigned to each questionnaire on page 1, and the cover sheet will be torn off before the data is tabulated. The code numbers assure the confidential treatment of your answers. Your candid answers will be appreciated and are necessary for the success of the study.

Thank you for your cooperation.

Name:____

Code Number______

(I)

In the ten spaces provided below please make 10 different statements in response to the question: "Who am I?" Give these answers as if you were giving them to yourself, not to somebody else. Write fairly rapidly, giving the first 10 responses to the question that come to your mind.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

(II)

In the ten spaces provided below please make 10 different statements in response to the question: "Who was I?" Give these answers as if you were giving them to yourself, not to somebody else. Write fairly rapidly, giving the first 10 responses to the question that come to your mind.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

(III)

In the ten spaces provided below please make 10 different statements in response to the question: "Who will I be?" Give these answers as if you were giving them to yourself, not to somebody else. Write fairly rapidly giving the first 10 responses to the question that come to your mind.

1.

2.

3.

4.

5.

6.

7.

8.

9.

10.

Incomplete Sentences Blank - Form I

Complete these sentences to express your real feelings. Try to do every one. Be sure to make a complete sentence.

1. If I could change my name ______

2. If you trust most people, you are apt to get ______

3. My goals are ______

4. When I find it difficult to carry out a job, I ______

5. When I compare the kind of person I am now with the kind of person I use to be, I ______

6. Intimate relationships for me are ______

7. I am sure that people think of me as ______

8. For me to really be a person ______

9. A person who has lost his wallet, if he has his name and address in it, can probably expect ______

10. The things that I want out of life are ______

11. Whenever there is more work to be done than I anticipated, I ______

14

12. When new fashions and ideas come along, I____________________

13. In the presence of women I feel____________________

14. The opinions of others are____________________

15. When somebody confuses me with someone else, I____________________

16. When you are having difficulties, you can expect other people to____________________

17. I am looking for____________________

18. Sometimes I get so preoccupied with some activity____________________

19. The person I was yesterday and the person I am today are____________________

20. At small parties, I feel____________________

21. When I think about the kind of person I appear to be to others, I____________________

22. I am not myself, when____________________

23. Without law enforcement, most people would____________________

24. When I am 30, I expect to be____________________

25. When I have a required task, I____________________

26. The things that once were important to me____________________

27. Working very closely with another person____________________

28. I feel that most people who meet me for the first time think I am____________________

29. I feel like an imposter, when____________________

30. If you don't know much about something, you can expect other people to____________________

31. I feel committed to____________________

32. Knowing how to do a job well____________________

33. What I think changes most about a person____________________

34. To be aloof and inaccessible____________________

35. Proving myself to others is____________________

36. Knowing who I am is____________________

37. I think the person who trusts in his own abilities____________________

38. My life is build around____________________

____

39. When I stop getting immediate results from what I am working on

____

40. From past experience I have learned that I____________________

____

41. When I have to be by myself for any length of time____________________

____

42. When I walk into a room and know that everybody is looking at me, I____________________

____

43. Pretending to be somebody you aren't is____________________

____

44. The man who invests his efforts in the future____________________

____

45. In making plans for the future, I____________________

____

46. I feel most inferior with regard to my abilities, when I____________________

____

47. If it were possible to go back in time and see myself as I used to be, I would probably feel that I____________________

____

48. When I am separated from a person I love____________________

____

49. When someone tells me something about myself, I____________________

____

50. Knowing where a person belongs is____________________

____

51. The best policy when someone offers you something for nothing is ____________________

52. I anticipate ____________________

____

53. I think that the anticipation of achievement ____________________

____

54. When I try to make sense out of my past life ____________________

____

55. I feel that friendship is ____________________

____

56. In a comparison of the way I see myself and the way others see me, I think ____________________

____

Study of Opinions

Form A

A number of statements about important social and personal questions are given below. Indicate your own personal opinion by circling the one alternative that best represents how much you agree or disagree with each statement. This is not a test in which there are "right" and "wrong" answers. People have different reactions to these statements. What is wanted here is your own personal reaction. Please give some answer to each question.

A = Agree strongly
a = Agree somewhat
? = Undecided
d = Disagree somewhat
D = Disagree strongly

A a ? d D 1. Life is chaotic, without direction or meaning.

A a ? d D 2. There is great pleasure in giving a job all you've got.

A a ? d D 3. The artist and professor are probably more important to society than the business man or manufacturer.

A a ? d D 4. It isn't necessary to be a chameleon and be all things to all people in order to get ahead in life.

A a ? d D 5. Many people don't have what it takes to finish a job.

A a ? d D 6. We arrive at wisdom through our intimacies with other people.

A a ? d D 7. Familiarity breeds contempt.

A a ? d D 8. The major decisions a person makes are guided by the plans he has for the future.

A a ? d D 9. Most of our social problems would be solved if we could somehow get rid of the immoral, crooked and feeble-minded people.

A a ? d D 10. The worst thing that can happen to a person is to turn out to be a nobody.

A a ? d D 11. An urge to jump from high places is probably the result of unhappy personal experiences rather than something inborn.

A a ? d D 12. If given a chance, most people would take advantage o you.

A a ? d D 13. One of the most important things children should learn is when to disobey authorities.

A a ? d D 14. A close physical relationship in marriage plays an important part in one's happiness.

A a ? d D 15. A person doesn't change much once he has started out in the world.

A a ? d D 16. People wouldn't do any work at all if they didn't have to.

A a ? d D 17. It is best not to become very intimate with another person, because it always leads to frustration and disappointment.

A a ? d D 18. One doesn't depend on others to tell him who he is and where he is going.

A a ? d D 19. The man who walks in solitude is the one who experiences the greatest enjoyments of which nature is capable.

A a ? d D 20. Although there is a lot of talk nowadays about the loss of individuality in our mass society, most of us feel that we are distinct individuals in our own right.

A a ? d D 21. We all need the love of those who are close to us.

A a ? d D 22. An individual should adjust his thinking about himself to concur with the views of the majority.

A a ? d D 23. The most difficult task for a person is to decide what kind of life he wants.

A a ? d D 24. Books and movies ought to give a more realistic picture of life even if they show that evil sometimes triumphs over good.

A a ? d D 25. Even if it were possible to start life all over again, most of us would decide to be the same person we are.

A a ? d D 26. If you are not getting results on a job, it is best to quit before you expend too much energy.

A a ? d D 27. In spite of what you read about the wild sex life of people in important places, the real story is about the same in any group of people.

A a ? d D 28. A man's present life is continuous with his past.

A a ? d D 29. A person can be confident of getting recognition from those who count.

A a ? d D 30. No sane, normal, decent person could ever think of hurting a close friend or relative.

A a ? d D 31. People ought to pay more attention to new ideas, even if they seem to go against the American way of life.

A a ? d D 32. Life is guided by certain principles one learns as a child.

A a ? d D 33. What really matters is what other people think; it is not enough just to be sure of oneself.

A a ? d D 34. What the youth needs most is strict discipline, rugged determination, and the will to work and fight for family and country.

A a ? d D 35. A man can place his trust in the future.

A a ? d D 36. Most of us are not pleased with the "type" of person we appear to be to others.

A a ? d D 37. If it weren't for the rebellious ideas of youth there would be less progress in the world.

A a ? d D 38. It is foolish to worry about your own capabilities, because most of us can be trusted to come through when it really counts.

A a ? d D 39. The findings of science may some day show that many of our most cherished beliefs are wrong.

A a ? d D 40. Most people can plan their lives so that they get what they want.

A a ? d D 41. In times like these, no one can trust what the future will bring.

A a ? d D 42. Sex crimes such as rape and attacks on children, deserve more then mere imprisonment, such criminals ought to be publicly whipped or worse.

A a ? d D 43. At times, nearly everyone would like to be someone else.

A a ? d D 44. Every person should have complete faith in some supernatural power whose decisions he obeys without question.

A a ? d D 45. Most young people today feel it is foolish to do more work than is necessary to "get by."

A a ? d D 46. Being too involved with someone else takes something away from yourself.

A a ? d D 47. The only thing to trust is mistrust.

A a ? d D 48. Most people act and express themselves quite differently at different times.

A a ? d D 49. It is better to have a few close friends than many acquaintances.

A a ? d D 50. A man can be sure of himself regardless of the opinions of others.

A a ? d D 51. Human nature being what it is, there will always be war and conflict.

A a ? d D 52. Most of us have a fairly good idea what we'll be ten years from now.

A a ? d D 53. It is highly unlikely that astrology will ever be able to explain anything.

A a ? d D 54. It isn't necessary to be always proving yourself to others, it is enough to know it yourself.

A a ? d D 55. Distant goals deserve more effort than momentary pleasures.

A a ? d D 56. A person's own estimates of his capabilities are likely to change quite frequently.

A a ? d D 57. Most honest people admit to themselves that they have sometimes hated their parents.

A a ? d D 58. It is rarely safe to trust a stranger.

A a ? d D 59. A man's tastes and sentiments remain relatively constant.

A a ? d D 60. No weakness or difficulty can hold us back if we have enough will power.

A a ? d D 61. No one is the same person from day to day.

A a ? d D 62. A person who can do a job doesn't have to feel inferior about anything.

A a ? d D 63. The present has meaning in terms of the future.

A a ? d D 64. What a bore it is, waking up in the morning always the same person.

A a ? d D 65. Most people don't know where they are headed.

A a ? d D 66. When you come right down to it, most people will help you when you are in trouble.

A a ? d D 67. A man should work hard when he works.

A a ? d D 68. The older you get, the more you realize that you have to trust other people.

A a ? d D 69. Most people know who they are and where they belong.

A a ? d D 70. Every man has felt he would like to cut his moorings and quit his home, his parents, and his friends.

A a ? d D 71. Most people don't realize how much our lives are controlled by plots hatched in secret places.

A a ? d D 72. It is the exceptional person who plans for the future.

A a ? d D 73. Nowadays more and more people are prying into matters that should remain personal and private.

A a ? d D 74. Even when they are quite alone, most people worry a great deal about what others think of them.

A a ? d D 75. It is only human to worry a lot about finding yourself.

A a ? d D 76. It is essential for learning or effective work that our teachers or bosses outline in detail what is to be done and exactly how to do it.

Autobiographical Questionnaire

Please answer the following questions as frankly and accurately as possible. They are designed to describe aspects of your background and life history. All information is confidential and for research purposes only.

1. School or college:__________________________ Class:___________

2. Your age (to nearest birthday)__________years Sex:________

3. What is your religious preference? (Check one)

   Catholic ( ) Jewish ( ) Protestant ( ) (specify)__________

   None ( ) Other ( ) (specify)________________________________

4. What is your racial background? (Check one)

   White ( ) Negro ( ) Oriental ( ) Other ( ) (specify)________

5. In which of the following geographical areas did you spend most of your time during childhood? (Check one)

   Farm ( ) Town ( ) City ( ) Suburb ( )

6. Do you or did you have any brothers or sisters? Yes ( ) No ( )

   If yes, please list them in order of birth (include deceased brothers and sisters). Be certain to include yourself in the appropriate position. Then fill in the requested information in Columns A-E. (Columns F and G will be used in answering questions 7 and 8.)

A	B	C	D	E	F	G
Sex	Year of Birth (approx.)	Education Highest grade or degree	Occupation	If dead, Age at death	M	F

7. In Column F (Question 6) rank order the children in your family including yourself from the one most favored by your mother (rank 1) to the one least favored by her.

8. In Column G (Question 6) rank order the children in your family including yourself from the one most favored by your father (rank 1) to the one least favored by him.

9. What is/was your father's major occupation?____________________

Did/does your mother have a major occupation (other than housewife)?

(a) Yes ( ) Specify________________________(b) No ( )

10. What formal education did your parents have? (Please circle highest grade completed.)

Father: Grade 1 2 3 4 5 6 7 8 High School 1 2 3 4

College 1 2 3 4; Post-graduate 1 2 3 4+

Highest degree held:_______________________

Mother: Grade 1 2 3 4 5 6 7 8 High School 1 2 3 4

College 1 2 3 4; Post-graduate 1 2 3 4+

Highest degree held:_______________________

11. What are the best things about your mother's personality?__________

12. What are her personality weaknesses?__________________________

13. What are the best things about your father's personality?__________

14. What are his personality weaknesses?__________________________

15. What are the best things about your personality?________________

16. What are your personality weaknesses?_________________________

17. How would you describe your relationships with your parents when you were a child?____

18. How have your relationships with your parents changed since you were a child?____

19. During childhood and later on did you have a particularly close relationship with any of your brothers or sisters? Yes ( ) No ( )

If Yes:
(a) Which brother or sister? (Place a check next to his position in the birth order chart of Question 6.)
(b) What was the nature of the relationship?____________________
____

20. Relationships in Your Family:

Put a check in the appropriate spaces below. Please answer every question.

	Father	Mother	A Brother	A Sister	Other (Specify)
Who made most of the decisions?	( )	( )	( )	( )	( )
Who has taken the most interest in you?	( )	( )	( )	( )	( )
Whom do you most respect?	( )	( )	( )	( )	( )
Whom are you closest to now?	( )	( )	( )	( )	( )
Who would you most like to be like?	( )	( )	( )	( )	( )
Who best understands you.	( )	( )	( )	( )	( )
Whose praise do you seek most?	( )	( )	( )	( )	( )
Whom did you side with in quarrels?	( )	( )	( )	( )	( )
Who was the most jealous?	( )	( )	( )	( )	( )
Who did you shy away from the most?	( )	( )	( )	( )	( )

21. Did you have any unusual accidents or illnesses during childhood or later on? Yes ( ) No ( )

If Yes:
(a) What were they?____________________
(b) How old were you at the time?__________years.
(c) Did you have to be cared for away from home? Yes ( ) No ( )
If Yes: How long?__________

22. Did you have any unusual disabilities in childhood or later on?
Yes ( ) No ( )

If Yes:
(a) What were they?____________________
(b) How old were you at the time?__________years.
(c) Did you have to be cared for away from home? Yes ( ) No ( )
If Yes: How long?__________

23. What is the most shocking or traumatic incident during childhood that you can recall? State the age at which the experience occurred and describe it briefly. Age:___________years.

Description of event as you remember it:_________________________

____

____

24. How would you rate your childhood?

Very happy	Happy	Average	Unhappy	Very unhappy
( ) 1	( )2	( )3	( )4	( )5

25. Indicate which of the following things have caused you unhappiness during the last few months. Mark each one either 0, 1, or 2.

0 = has not caused unhappiness
1 = has caused some unhappiness
2 = has caused a good deal of unhappiness

____your personal appearance
____your health
____lack of a real future
____that you feel very anxious
____difficulties with the opposite sex
____that you can't concentrate on required tasks
____lack of a sense of belonging
____loneliness
____too much competiveness from others
____troubles with your family
____troubles with your friends
____troubles with your teachers
____difficulties in planning your career
____that you can't fit together all the different things you do

26. Below are listed a number of activities that may provide satisfactions in a person's life. Indicate the relative importance of each of these activities to you by assigning it a rank from 1 to 7. (Write "1" next to the activity which you consider to be of the greatest importance in providing satisfaction in life; write "2" next to the one of next greatest importance, and so on until each activity has been given a rank.)

____ Career or occupation
____ Family relationships
____ Maintaining the right social standing
____ Religious and church activities
____ Leisure-time recreational activities
____ Participation in politics
____ Participation in activities directed toward the betterment of others

27. Group Memberships

In the spaces below list the various groups of which you are a member. (Include fraternal organizations, young people's religious groups, professional and honorary societies, campus clubs, athletic teams, etc.) Indicate your degree of participation in each of these groups by placing a check in the appropriate spaces.

	Amount of Participation				
	Very much	Much	Average	Little	None
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )
________________	( )	( )	( )	( )	( )

28. How important is it for you to have plans for the future rather clearly known in advance? (Check one)

Very important	Fairly important	Not very important	Very unimportant
( )1	( )2	( )3	( )4

29. How important to you, personally, is it to get ahead in life? (Check one)

Very important	Fairly important	Not very important	Very unimportant
( )1	( )2	( )3	( )4

30. If you had your choice, which of the following would you most like to be? (Check one)

( ) Independent
( ) Successful
( ) Well liked

31. What is your major field in college? ____________________

32. Have you shifted majors? What was your previous major fields?

____

Career or Occupation

33. Have you chosen your career or occupation? Yes ( ) No ( )

If Yes:

(a) Please name your anticipated career field: ______________

(b) What were the major reasons for your choosing this particular occupation?

(1) ______________________________

(2) ______________________________

(c) In the space below, briefly outline the major steps you see yourself following in order to achieve your vocational goals.

(d) How old were you when you first began to seriously consider this as your occupational choice? ____________ years.

(e) How committed do you feel to this particular career choice?

Very committed	Committed	Uncommitted	Very uncommitted
( )1	( )2	( )3	( )4

(f) What other fields of employment have you seriously considered? ______________________________

(g) How likely is it that you may give up the present choice in favor of one of these other choices or some new choice?

Very unlikely	Unlikely	Likely	Very likely
( )1	( )2	( )3	( )4

If No:

(a) Since you have not chosen your long-run career or occupation, how concerned do you feel about not having decided?

Very unconcerned	Unconcerned	Concerned	Very concerned
( )1	( )2	( )3	( )4

(b) How much time and thought do you give to the consideration of various careers or occupations that you might pursue?

None	A Little	Some	Much	Very much
( )1	( )2	( )3	( )4	( )5

A Fortran-IV Program for Evaluating Internal Consistency and Single-Factoredness in Sets of Multilevel Attitude Items[1]

Donald G. Morrison, Donald T. Campbell,

and Leroy Wolins

Northwestern University and Iowa State University

Distributed by the Vogelback Computing Center, Northwestern University, Evanston, Ill. Program NU C0048 ATTANAL

---

[1] Supported in part by U.S. Office of Education, Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

A Fortran-IV Program for Evaluating Internal Consistency and Single-Factoredness in Sets of Multilevel Attitude Items

Donald G. Morrison, Donald T. Campbell,
and Leroy Wolins

When we summarize a person's responses to a number of questions by a single score, and when we give that score a content-referent name, we assume that the questions and answers are all tapping a common trait, attitude, or factor. The appropriateness of this assumption obviously needs checking, and procedures for so doing have been offered at every stage of the history of attitude testing, starting with the almost-never-used coefficient of irrelevance of Thurstone and Chave (1929).

Currently, the dominant mode of stating this requirement is in terms of Guttman's (1944, 1950) scale analysis procedures, where the requirement becomes one of "unidimensionality," and the degree of this measured by a "reproducibility coefficient." This tradition is deemed thoroughly inadequate inspite of an excellent clarification of the ideal condition. This is not only because the 90% reproducibility requirement is arbitrary, nor only because chance levels of reproducibility are almost never reported and may be as high as 80%, but also because even the very best modifications and provisions for tests of significance (Sagi, 1959; Goodman, 1959) test only the item-set's departure from total item independence and would pass as unidimensional a set of items in which two subsets were highly related within subset but totally independent between subsets.

In the end, the more adequate theoretical models seem to come from the tradition of psychometric test analysis, including factor analysis. The requirements become high internal consistency reliability and single factoredness. While this tradition leaves much to be desired, it is judged to be much more appropriate and usable than the Guttman Scale Analysis tradition. Further, the Guttman requirement of unidimensionality translates into a psychometric requirement of single factoredness (Loevinger, 1948; Siegmann, 1954; Lumsden, 1961).

Both the psychometric tradition in scoring items as pass-fail, and the Guttman tradition as practiced, deal with dichotomously scored items. For some conceptual purposes these are preferable to multilevel responses (such as: Agree Strongly, Agree, Uncertain, Disagree, Disagree Strongly). But for analysis purposes they present many problems--due perhaps mainly to the lack of appropriate mathematical models. These are illustrated by ceilings less than 1.00 for intercorrelations among items where marginals are uneven. (For example, if one item is passed by 90% and another by but 10% the maximum positive point correlation or phi coefficient is .11). These produce irrelevant "difficulty factors" (Wherry and Gaylord, 1944). A similar ceiling is placed upon Kuder-Richardson type reliability coefficients (Horst, 1953). Multilevel scores do not entirely eliminate these problems (Carroll, 1961) but they reduce it to trivial proportions.

This program is designed to operate on data cards containing one-digit scores for each item (e.g., 1 through 5) which are converted from the original answering language into a scoring language, so that larger scores equal more of the attitude in question. It is assumed that the items will be of two types, and some analyses will be made of the two types separately, to answer questions as to response-set or direction-of-wording effects. Thus in many scales, half of the items would be worded so that agreeing represented pro attitudes, half so that agreeing represented con attitudes. In the program these two types would be designated A and B or A positive and B positive. This particular designation emerges from our recommended best practice, the "polar choice" format. Each item is a choice between a pair of statements, often dealing with the same illustrative content, one positively advocating the pro position, the other the con. This eliminates that ambiguity of the "disagree" response in the usual format which is due to the fact that one can disagree with a statement for so many different reasons. Instead, one has to locate his attitude on a dimension anchored by a positive statement at each pole. (This format should not be called "forced choice" since in the established meaning of that method, the choice is between or among items coming from two or more dimensions, equated for popularity. Here, the two statements are poles of a single dimension.) Several investigators have independently hit upon this format (e.g., Loevinger, 1962; Owens, 1947, Kassarjian, 1962; Cook & Selltiz, 1964; Berkowitz & Wolkon, 1964). Multilevel responses are achieved by the following instructions:

CHOICES OF OPINION

This is a survey of attitudes and opinions on a variety of topics. Each item consists of two alternatives, A and B, between which you are asked to choose by circling one of these indicators:

A - Statement A is entirely preferred to Statement B as an expression of my opinion.

a - Statement A is somewhat preferred over Statement B.

? - I cannot choose between A and B.

b - Statement B is somewhat preferred to Statement A.

B - Statement B is entirely preferred to Statement A as an expression of my opinion.

Please show your attitude leanings on each item, even though you do not feel strongly on the topic or do not feel well informed. Please choose between alternatives, even though both may seem acceptable to you, or both unacceptable.

The questions are then presented as follows, (taking a few items from a student's scale):

04 Capital Punishment

A a ? b B
1. A. Crimes of violence are rarer in areas where capital punishment is employed.
B. The death penalty has no effect on the rates of violent crimes.

A a ? b B
2. A. Most vicious criminals, if given the proper opportunities and guidance, will be rehabilitated and make contributions to society.
B. Most vicious criminals are incapable of change, and thus pose such a threat to society that they should be executed.

A a ? b B
3. A. Capital punishment is actually more humane than is imprisonment for the rest of one's life.
B. The compassionate judge sentences a man to life imprisonment, when he could be condemned to death.

A a ? b B
4. A. Abolishment of the death penalty would make no appreciable effect on whether or not prisons are overcrowded.
B. Penal institutions would soon become seriously overcrowded if capital punishment were to be abolished.

In this illustration, items 1 and 3 are "A-positive" in that alternative A is the more favorable to capital punishment. For these, a scoring key might translate A, a, ?, b, B into 5, 4, 3, 2, 1 respectively. Items 2 and 4 are "B-positive," and A, a, ?, b, B would correspondingly be scored 1, 2, 3, 4, 5.

Which ever format is used, the program calls for two-digit sums of the A-positive items, the same for the B-positive, and a total of all items.

Reliability

The program reads out for each item and total the means and standard deviations, and a table of all intercorrelations among the items and sums. (Where totals are involved, these are uncorrected for part-whole effects.) For reliability, the program first takes the correlation between A-positive and B-positive totals and using it as a split-half coefficient, applies a Spearman-Brown correction to it (Guilford, 1954, p. 354). This is labeled in the read-out as "split-half reliability," although this is somewhat misleading, for if there is a direction-of-wording effect, this is probably the most dissimilar of all possible splits, the lowest of possible split halves. The best indicator of reliability is the coefficient labeled "Kuder-Richardson for total." This is the modern internal consistency reliability independently invented by Kuder, Richardson, Hoyt, Guttman, etc.(Guilford, 1954, pp. 380-385). In the form used here for multilevel scores it is what Cronback (1951) calls "Coefficient Alpha," item variances taking the place of the usual pq term.

$$\text{Reliability} = (\frac{n}{n-1})(1 - \frac{\Sigma\sigma_i^2}{\sigma_t^2})$$

where n = number of items, σ_i^2 = variance of an item, σ_t^2 = variance of the total score.

Reliabilities for scales constructed on novel topics are often so low that one can ask whether or not they significantly differ from zero. Hoyt's

(1941) approach to the coefficient derives it from an analysis of variance of the main effect of persons tested against the persons-times-items interaction. The relation between this F and the reliability is $r_{k_r} = 1 - 1/F$. The degrees of freedom are (respondents - 1) and (respondents - 1) x (items -1). If we have 100 persons and 8 items, the degrees of freedom would be 99 and 693. For this, an F table shows an F of approximately 1.40, significant at the .01 level. This translates to a reliability of .29. The more persons and the more items, the lower the reliability value required for significance. (These significance manipulations are not in the computer read-out at the present time.)

Reliability in the above senses is a function of both internal consistency and test length. To produce an index independent of test length, it is suggested (e.g., Guilford, 1954, p. 387) that these values be converted to the inferred average inter-item correlation, or the reliability of a one-item version of the test. The formula is a variant on the Spearman-Brown correction formula:

$$r_{ij} = \frac{r_{tt}}{n-(n-1)r_{tt}}$$

where n equals the number of items, r_{tt} the reliability.

One must get used to low values if this index is to become meaningful. A value of .20 is regarded as very good for an intelligence test item. Student researches at Northwestern show values ranging from .03 to .30. The published values of the famous Ethnocentrism scale (Adorno et al., 1950) show a reliability of .91 for 34 items. This translates as an average inter-item correlation of .23. For the F scale of authoritarian personality trends they report as typical a value of .90 for 45 items, which translates to an r_{ij} of .20. In our own work at Northwestern, the E scale holds to approximately the level they found, but F scale values are much lower, .10 being typical (Campbell, 1965), even when only original F items are used. Correcting for direction-of-wording effects by having half of the items such that the democratic or nonauthoritarian might agree with them reduces the value of r_{ij} to less than .05. In general, the E scale is typical of psychology's successful attitude scales, and a .20 r_{ij} value must be regarded as highly successful. Note that the program computes the actual average coefficient (as TOTAL AVE).

Response Set

If there are direction-of-wording factors in the usual agree-disagree single item format (Cronbach, 1946; 1950; Chapman & Bock, 1958; Chapman & Campbell, 1959; Campbell, 1965), or response sets such as might result from primacy or recency effects in the A-B polar choice format, then these would make the intercorrelations among items unhomogeneous, raising correlations where both items were worded in the same direction, lowering them where they differ in direction. To examine for this, the correlations have been averaged by type, A with A, B with B, and A with B. If the A-A correlations and B-B correlations average higher than the A-B ones, response set or direction-of-wording effects are indicated.

Item-Total Correlation

A minimum requirement for a scale is that each item correlate with the sum of the other items positively, or to a significant degree. The item-total correlations from the initial read-out do not offer this information, as they are inflated by the identity of each item with its own contribution to the total. For this reason, the item-total coefficients have been corrected to remove this self correlation or part whole component. The formula used is this:

$$r_{it_{Cor.}} = \frac{r_{it}\,\sigma_t - \sigma_i}{\sigma_t^2 + \sigma_i^2 - 2r_{it}\,\sigma_t\,\sigma_i}$$

It always makes the correlation smaller, more so the larger the variance of the item. These values are read out, and along with them a record of the direction of wording of each item (put in on program control card).

Testing the Hypothesis of Single-factoredness

We have already performed some minimal tests of cohesiveness: If our A+ and B+ halves correlate significantly higher than zero, if our Kuder-Richardson reliability is significantly greater than zero, if each item correlates with the sum of the other items significantly greater than zero, then certain minimal standards may be regarded as having been met: there is significant cohesiveness among the items. Similarly, if we had applied a Goodman test to the Guttman reproducibility and found it significant, then some significant degree of cohesiveness among the items would have been demonstrated. But all of these criteria could have been met by patterns of relationships which are far from single-factored. Suppose, for example, that four of our eight items (two A+, and two B+) correlated with each other at the r = .80 level, and not at all with the other four items, while the other four items correlated highly too among themselves. This would be a clear cut instance of a two-factored set of items, yet it would meet all of our criteria of minimal cohesiveness. (It must be emphasized that Guttman scale analysis would also be "fooled" by such a two-factored set of items with exceptional cohesiveness among the parts.)

A look directly at the intercorrelations among the items will help. If, in fact, all of the items correlate with all of the other individual items positively, this is encouraging. If an item that has low correlations with one other item tends also to have low correlations with other items, this is further confirmation (Guilford, 1954, p. 473, criterion of proportionality). To test the matter more precisely, we must go to factor analysis. But it will not do to feed the data cards into a standard computer program for factor analysis. These programs are designed for "passive," "descriptive" factor analysis, and are systematically biased in the direction of extracting too many factors (the fewer cases the correlations are based upon, the more factors they will extract, multiplying unneeded entities in a most unparisimonious way) and in the direction of distributing the variance too evenly over the factors extracted, against a strong general factor. (Guilford, 1954, p. 500, paragraph 4, is 180° wrong.) What we need is a modification of factor analysis in the direction of a test for single-factoredness.

What we first present is a modification of Thurstone's Centroid Method, designed to be optimal if single-factoredness were actually the case. In such a case, the square of the first-factor loading is equal to the communality, there being no other factors. We modify Guilford's presentation of first factor extraction (p. 487-489) to iterate for the first-factor loadings optimal under this condition. Iteration is a trial-and-error procedure which approaches an optimal value by successively closer approximations. As an initial estimate of communality, we use for each item the square of the item-total-corrected-correlation (though because of the iteration procedure, any initial value will give the same result, eventually). We do not write this in the diagonal of the matrix, however, for reasons of later convenience. The sum that Guilford(Table 16.7, p. 488) calls E we do in two components. First we sum the coefficients in each column with the diagonal value left blank (called SUM on the computer read-out), then write below this sum the communality estimate (= h^2 in Guilford, = H SQ in the computer read-out), then add these two (the Sum and the h^2 estimate) to get the value Guilford calls E. Solve for a_1 as Guilford indicates. This is the first-factor loading for each item, first estimate. But if there is but one factor, then the communality is equal to a_1^2. This will differ slightly from the communality estimate we used previously, so let's compute E over again, using these new values (the SUM component of course remains the same). We continue this process until the values for a_1 cease to change. If some are negative, an additional rule needs to be invoked. Whenever a sum is negative, substitute 0.000 for it, i.e., no negative factor loadings are allowed on the initially hypothesized single-factor set (i.e., one allows an item to be irrelevant, but not to be so misguided as to be scored in the wrong direction).

This mode of extracting one factor is included because of its appropriateness as a teaching device, for desk calculator computation. For pedogogical purposes, the matrix of inter-item correlations predicted by these factor loadings are printed out, and then a residual matrix representing the difference between these values and the original correlations (presented in the second sheet of the read-out). The mean and sigma of these residuals is also reported. (This sigma is of a special sort--the root mean squared deviation around zero, not around the mean residual, if these differ $\sqrt{\frac{\Sigma r_{es}^2}{n}}$.)

How large would the sigma of residuals be if we did have a truly single-factored case? It would be larger the smaller the number of cases on which the coefficients were computed. It would be smaller the higher the general level of initial correlations. Let us consider a hypothetical monte-carlo study run on the computer, in which we combine normal random numbers so that the "true" factor loadings are all .71, and the expected correlations are all .50. If we did such a study with a small number of monte-carlo "persons," such as 25, we would find that many correlations deviated quite widely from the true value of .50. There would be some as high as .80 and other as low as .20, and some even more extreme. If, however, the number of persons were 100, then the range would be more likely from .35 to .65. The formula for this expectation is that for the standard deviation of a correlation coefficient:

$$\sigma_r = \frac{1-r^2}{\sqrt{N-1}}$$

We have experimented with using this as a single basis for determining the expected sigma of residuals if only one factor were present. This effort has failed because in a monte-carlo study such as described, the obtained factor loadings fit the sample correlations better than the true ones would, the variance of residuals as a result is much smaller than this formula estimates (particularly for small numbers of items). Lacking a good estimate for the expected variance of residuals under the single-factored case stops our desk-calculator approach at this point.

The centroid method of factoring is not the optimal model of single-factoredness. It distributes residuals evenly over initially high correlations and originally zero correlations. As we can see from the formula for r, larger residuals would be more expected from true r's of zero than from high ones. Lawley's Maximum Likelihood Method (Harmon, 1960, pp. 366-381) provides a single factor (or if asked, a multi-factor) model that distributes the residuals more optimally. It does this by starting from a better factor model than the Centroid, namely Principal Components, and then iterating to minimize a summed weighted residual.

For these residuals, there is a test of significance. One computes a X^2 according to this formula:

$$\text{Chi-Square} = N\Sigma \frac{\text{Residual Squared}}{(1-a_{1_1}^2)(1-a_{1_2}^2)}$$

One squares each residual, then divides it by the product of the non-communality of each of the items involved in the residual (i.e., the $(1-a_1^2)$ for each item). One sums all of these fragments, and multiples by the number of cases. This Chi-Square has degrees of freedom of $\frac{(n-m)^2 - n - m}{2}$ where n is the number of variables and m is the number of factors.

The computer read-out provides the factor loadings resulting, the residual matrix (below the diagonal) and the Lawley Chi square with the degrees of freedom. Note one modification of Lawley's maximum likelihood procedure which we have made: even though it sometimes results in larger chi squares, we have not allowed any factor loadings to go below zero, i.e., to become negative. Such a negative factor loading would be equivalent to telling us that we did not know how to score the item in the first place, that the scoring should have been reversed. Such an error we believe should contribute to the imperfection of the scale, and be signified by larger residuals and chi-square. Other aspects of the program have been borrowed from Haemmerle (1964) as indicated by the "IOWA" on the program listing.

Very often the Lawley chi square will be significant, indicating more than one factor, even for very good scales where all items correlate very highly with each other. Especially is this so for large numbers of cases, which make small residuals significant. For this reason, we sought an index of the degree to which the ideal of single-factoredness is approached which would be independent of the number of cases and descriptive of the degree of relationship. This index, called the Wolins Index in the computer read-out, has the following logic. It should be at 1.00 when the first factor leaves only the residuals that are expected in a one factor case. It should read 0 were the first factor to extract nothing. It will not duplicate the reliability or the average factor loading in indicating how big the first factor is--rather it will compare the first factor effect with the original level of the correlation matrix to be explained.• The formula employed is

$$I = \frac{\sigma_m - \sigma_o}{\sigma_m - \sigma_e}$$

where σm is the standard deviation of the original correlations from zero, representing the starting point or the variance to be explained; σo is the observed sigma of residuals after the Lawley maximum likelihood extraction of one factor, σe is the expected sigma of residuals if only one factor is involved. Computing σe is the most difficult problem. It has been derived from Lawley's work as detailed in the Appendix.

The utility of this index is as yet untested. We propose to develop experience with it by running through the program classic attitude scales that have become meaningful reference points in the literature. We also propose to simulate with random number generators attitude scales of known factorial composition. It is probable that the index will rarely go below .50 even for clearly multifactor scales--for the first factor always extracts at least one of the factors. A value of .75 may turn out to be the lower bounds of acceptable unidimensionality.

References

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. The authoritarian personality. New York: Harper, 1950.

Berkowitz, N. H., & Wolkon, G. H. A forced-choice form of the F scale, free of acquiescent response set. Sociometry, 1964, 27, 54-65.

Campbell, D. T. On the use of both pro and con items in attitude scales. Duplicated Research Report, Department of Psychology, Northwestern University, 1965.

Carroll, J. B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Chapman, L. J.,& Bock, R. D. Components of variance due to acquiescence and content in the F scale measure of authoritarianism. Psychological Bulletin, 1958, 55, 328-333.

Chapman, L. J., & Campbell, D. T. The effect of acquiescence response-set upon relationships among the F scale, ethnocentrism and intelligence, Sociometry, 1959, 22, 153-161.

Cook, S. W., & Selltiz, C. A multiple-indicator approach to attitude measurement. Psychological Bulletin, 1964, 62, 36-55.

Cronbach, L. J. Response sets and test validity. Educational and Psychological Measurement, 1946, 6, 475-494, also 1950, 10, 3-31.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Goodman, L. A. Simple statistical methods for scalogram analysis. Psychometrika, 1959, 24, 29-44.

Guilford, J. P. Psychometric methods. Rev. Ed. New York: McGraw-Hill, 1954.

Guttman, L. A basis for scaling qualitative data. American Sociological Review, 1944, 9, 139-150.

Haemmerle, W. J. Multiple method factor analysis. Statistical Laboratory, Iowa State University, February 3, 1964.

Harmon, H. H. Modern factor analysis. Chicago: Univer. Chicago Press, 1960.

Horst, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. Psychological Bulletin, 1953, 50, 371-374.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Kassarjian, W. M. A study of Riesman's theory of social character. Sociometry, 1962, 25, 213-230.

Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh, Vol. LX, 1939-40, Part I, pp. 64-82.

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychological Bulletin, 1948, 45, 507-530.

Loevinger, J. Measuring personality patterns of women. Genetic Psychology Monographs, 1962, 65, 53-136.

Lumsden, J. The construction of unidimensional tests. Psychological Bulletin, 1961, 58, 122-131.

Owens, W. A. Item form and "false-positive" response on a neurotic inventory. Journal of Clinical Psychology, 1947, 3, 264-269.

Sagi, P. C. A statistical test for the significance of a coefficient of reproducibility. Psychometrika, 1959, 24, 19-28.

Siegmann, P. J. A comparison of factor analysis with Guttman's scaling technique. Unpubl. Ph.D. diss., Ohio State Univ., 1954.

Thurstone, L. L., & Chave, E. J. The measurement of attitude. U. Chicago Press, 1929.

Wherry, R. J., & Gaylord, R. H. Factor pattern of test items as a function of the correlation coefficient: Content, difficulty, and constant error factors. Psychometrika, 1944, 9, 237-244.

Appendix A

A Derivation of the Expected Variance of Residuals

Leroy Wolins

The purpose of this report is to present a derivation of the expected variance of residuals formed by subtracting the inner product of a factor matrix from the correlation matrix from which the factors were derived. This expected variance of residuals assumes that the factors were extracted by means of Lawley's maximum likelihood method and depends upon the assumptions of that method.

Equation 1 presents the conventional chi-square test for the null hypothesis that m factors are adequate to explain the intercorrelations among n variables where the intercorrelations are based upon N observations. In (1),[*] $\bar{r}_{jk}$ represents the residual between the variables represented by row j and column k. The terms a_j and a_k are the specific variances of variables j and k, respectively and are computed by subtracting from unity the sum of squares of the factor loadings of variables j and k, respectively. Equation 1 indicates that each residual be divided by the product of the squared values of the two specific variances of the two variables involved in the residual. These values are then summed for all values where j, the row indicator, is less than k, the column indicator. Thus the summation of the elements in an upper triangular matrix with zeroes in the diagonal is indicated.

The degrees of freedom for this chi-square statistic is indicated in (2).

The crux of this derivation is recognizing in (1) that the summation term is the first term of the covariance between the squared values of the residuals and the product of the reciprocals of the squared specific variances of the two variables which involve the residuals. That is, one may imagine computing a covariance (or a correlation) based on pairs of "observations" as follows:

[*] The formulas referred to by (1) etc. are contained on pages 5 and 6 of this appendix.

"Observation"	$\bar{r}_{jk}$	$a_j^{-2}\ a_k^{-2}$
1, 2	$\bar{r}_{12}$	$a_1^{-2}\ a_2^{-2}$
1, 3	$\bar{r}_{13}$	$a_1^{-2}\ a_3^{-2}$
1, 4	$\bar{r}_{14}$	$a_1^{-2}\ a_4^{-2}$
$\vdots$	$\vdots$	$\vdots$
1, n	$\bar{r}_{1n}$	$a_1^{-2}\ a_n^{-2}$
2,3	$\bar{r}_{23}$	$a_2^{-2}\ a_3^{-2}$
2, 4	$\bar{r}_{24}$	$a_2^{-2}\ a_4^{-2}$
2, 5	$\bar{r}_{25}$	$a_2^{-2}\ a_5^{-2}$
$\vdots$	$\vdots$	$\vdots$
2n	$\bar{r}_{2n}$	$a_2^{-2}\ a_n^{-2}$
$\vdots$	$\vdots$	$\vdots$
(n-1)n	$\bar{r}_{(n-1)n}$	$a_{(n-1)}^{-2}\ a_n^{-2}$

Equation (3) indicates the full expression of this covariance. Equation (4) indicates the expected variance of residuals when the deviations from zero are squared and summed. The deviations are taken from zero rather than from the mean residual computed from the data because we are attempting to derive the expected variance of the residuals under null conditions, when m factors are sufficient. Under null conditions the expected mean residual is zero. Since the expected mean is known, the proper denominator is $\frac{n(n-1)}{2}$ rather than $\left[\frac{n(n-1)}{2} - 1\right]$ since we do not lose a degree of freedom by estimating the mean from the data. One may note that the variance of residuals, (4), is explicit in (3).

Equations (5) serve the purpose of putting another term in (3), $\sum_{j<k=1}^{n} a_j^{-2} a_k^{-2}$, in a more interpretable form. Equation (5) indicates that the sum of the elements of an upper triangular matrix with zeroes in the diagonal is equal to one-half of the sum of the elements in a square matrix minus the diagonal elements; which, in turn, is a function of the squared average of the square of the reciprocals of the specifics. That is, the term on the left of (5) will be large when the specifics are small on the average.

Equation 6 is formed by substituting in (3) equations (4) and (5) and by moving one term to the right of the equality sign with the appropriate change in sign.

Equation (7) states a well known fact that the expectation of a chi-square is approximately the degrees of freedom minus one-half when the degrees of freedom are greater than 30. A more accurate value for the expected value of chi-square can be obtained by referring to a chi-square table to the column labeled .50 and the row labeled with the degrees of freedom computed from (2). The table value should be used when the degrees of freedom computed from (2) is less than thirty.

Equation 8 is formed by substituting in (1), (6) on the right side of the equality sign and the expected value of chi-square on the left side of the equality sign, in this case (7). Equation 9 is formed by algebraic manipulation of (8) and is the solution for all m.

Since it is of special interest to develop $E(S^2_{\bar{r}_{jk}})$ when m = 1, (10) is derived from (9) for m = 1. However, (10), like (9), is predicated upon the degrees of freedom being greater than 30.

Equations (11) and (12) define again terms in (10).

Equations (13) through (20) show that the degrees of freedom will be greater than 30 when $m = 1$ and $n \geq 10$. If $n < 10$ and $m = 1$, (21) should be used instead of (10). Equation (21) may always be used but one may

$$E\left(S^2_{\bar{r}_{jk}}\right) = \left[\frac{E(\chi^2)}{\left[N - \frac{2}{3}n - \frac{5}{2}\right]} - C\right] \Bigg/ \left[\frac{n\overline{a^{-2}}\left(n\overline{a^{-2}} - 1\right)}{2}\right] \qquad (21)$$

find it difficult to locate tables of the chi-square distribution for degrees of freedom greater than thirty so that (10) may be convenient.

$$\chi^2 = \left[N - \frac{1}{6}(2n+5) - \frac{2}{3}m - 1\right] \sum_{j<k=1}^{n} \bar{r}^2_{jk}/a^2_j a^2_k \qquad (1)$$

$$df = \frac{1}{2}\left[(n-m)^2 - n - m\right] \qquad (2)$$

$$\sum_{j<k=1}^{n} \bar{r}^2_{jk} a_j^{-2} a_k^{-2} - \frac{\Sigma \bar{r}^2_{jk} \, \Sigma a_j^{-2} a_k^{-2}}{\frac{n(n-1)}{2}} = C_{\bar{r}^2_{jk}(a_j^{-2} a_k^{-2})} \qquad (3)$$

$$E(S^2_{\bar{r}_{jk}}) = E\left(\frac{\Sigma \bar{r}_{jk}}{n(n-1)/2}\right) , \quad \left(\begin{array}{l}\text{Under null conditions} \\ E(\Sigma \bar{r}_{jk}) = 0\end{array}\right) \qquad (4)$$

$$\sum_{j<k=1}^{n} a_j^{-2} a_k^{-2} = \frac{\sum_{j=1}^{n} \sum_{k=1}^{n} a_j^{-2} a_k^{-2} - \sum_{j=1}^{n} a_j^{-2}}{2} = \frac{n^2(\overline{a_j^{-2}})^2 - n\,\overline{a_j^{-2}}}{2} \qquad (5)$$

$$\Sigma \bar{r}^2_{jk} a_j^{-2} a_k^{-2} = E(S^2_{\bar{r}_{jk}}) \left[n^2(\overline{a^{-2}})^2 - n\,\overline{a^{-2}}\right] + C_{\bar{r}^2_{jk}(a_j^{-2} a_k^{-2})} \qquad (6)$$

$$E(\chi^2) \approx df - \frac{1}{2} \quad (\text{for } df > 30) \qquad (7)$$

$$E(x^2) = \frac{1}{2}\left[(n-m)^2 - n - m - 1\right] = \left[N - \frac{1}{6}(2n+5) - \frac{2}{3}m - 1\right]\left\{E(S^2_{\bar{r}_{jk}})\left[\frac{n^2(\overline{a^{-2}}) - n\,\overline{a^{-2}}}{2}\right] + C_{\bar{r}^2_{jk}(a_j^{-2} a_k^{-2})}\right\} \qquad (8)$$

$$E(S^2_{\bar{r}_{jk}}) \simeq \left[\left[\frac{\frac{1}{2}\left[(n-m)^2 - n - m - 1\right]}{\left[N - \frac{1}{6}(2n+5) - \frac{2}{3}m - 1\right]}\right] - C\right] \Bigg/ \left[\frac{n\,\overline{a^{-2}}(n\,\overline{a^{-2}} - 1)}{2}\right] \qquad (9)$$

$$\text{for } m = 1 \quad E(S^2_{\bar{r}_{jk}}) = \left[\frac{n^2 - 3n - 1}{\frac{1}{3}(6N-2n-15)} - C \right] \Big/ \left[\frac{n\,\overline{a^{-2}}(n\,\overline{a^{-2}}-1)}{2} \right] \tag{10}$$

N = sample size

n = number of variables

$$\overline{a^{-2}} = \frac{\frac{1}{a_1^2} + \frac{1}{a_2^2} + \frac{1}{a_3^2} + \ldots + \frac{1}{a_n^2}}{n} \tag{11}$$

$$C = \sum_{j > k=1}^{n} \bar{r}^2_{jk} \Big/ a_j^2\, a_k^2 - \frac{\Sigma\, r_{jk}\, \Sigma \frac{1}{a_j^2} \frac{1}{a_k^2}}{n(n-1) / 2} \tag{12}$$

$$\frac{1}{2}\left[(n-m)^2 - (n+m)\right] \overline{>} 30 \tag{13}$$

$$(n-1)^2 - (n+1) \overline{>} 60 \tag{14}$$

$$n^2 - 2n+1-n-1 \overline{>} 60 \tag{15}$$

$$n^2 - 3n \overline{>} 60 \tag{16}$$

$$n^2 - 3n - 60 \overline{>} 0 \tag{17}$$

$$n > \frac{3 \pm \sqrt{9 + 240}}{2} \tag{18}$$

$$n > \frac{3 + \sqrt{249}}{2} = \frac{3 + 16}{2} = \frac{19}{2} \tag{19}$$

$$n \overline{>} 10 \tag{20}$$

Appendix B

(to "A Fortran-IV Program for Evaluating Internal Consistency and Single-factoredness in Sets of Multilevel Attitude Items, by Donald G. Morrison, Donald T. Campbell, and Leroy Wolins.)

I. Description

This program calculates the various attitudinal testing statistics described above. Several separate sets of data can be run by following one set of control cards and data with another and so on for as many sets as desired.

Up to 47 items may be correlated and used in the analysis. There is no limit to the number of subjects. Each item in the scale must be classed as A positive or B positive for certain analyses.

II. Control Cards are to be punched as follows:

A. Problem Card

Cols.	Name	Contents
1- 2	ITEM	Number of items in scale
3- 5	NOSUBJ	Number of subjects
6- 7	NOXTR	Number of extra variables to enter correlation analysis only.
8-12	Blank	
	The following columns contain	
		1 if item is A positive
		0 if item is B positive
13	First item	
14	Second item	
15	Third item	
59	47th item (if used)	

B. Title Card

Columns 1 - 53 which may contain any alphanumeric information. This is used to title the output.

C. Variable Format Card

(for a more detailed description see Appendix on Variable Format)

The format card specifies the column locations of the data. It is best illustrated by example e.g., for a typical 8-item scale (10 X, 8 F 1.0, 2 F 2.0, F 3.0, 5 F 1.0) which means:

a. ship the first 10 columns (10 X)

b. take the 8 items from columns 11 - 18 (8 F 1.0)

c. A positive total in columns19 - 20
B positive total in columns21 - 22 (2 F 2.0)

d. Grand total in columns 23, 24, 25 (F 3.0)

e. Fine extra items in columns 26 - 30 (5 F 1.0)
to enter correlation analysis only.

III. Data Card Preparation

Data cards should be prepared in the following manner:

A. The user should use the first group of columns for identification of the subject and other information that would be useful to have with the data cards. This group of columns may be as large as is needed. The variable format card will contain instructions to skip these columns on reading the data cards. E.g., for a 10 column I-D, the Variable Format will begin (10 X. This information is not necessary so A may be absent.)

B. The items on the scale follow next and take up as many columns as necessary (in general ITEM columns).

C. The next three items are

1. The sum of the A positive scores
2. The sum of the B positive scores
3. The sum of all the scores for that subject.

D. Next are any other measurements that may be desired. These may be entered into the analysis of correlation, mean and standard deviation without disturbing the item analysis of the elements of part B.

IV. Card Order

I. ID cards and program deck

II. Control Cards

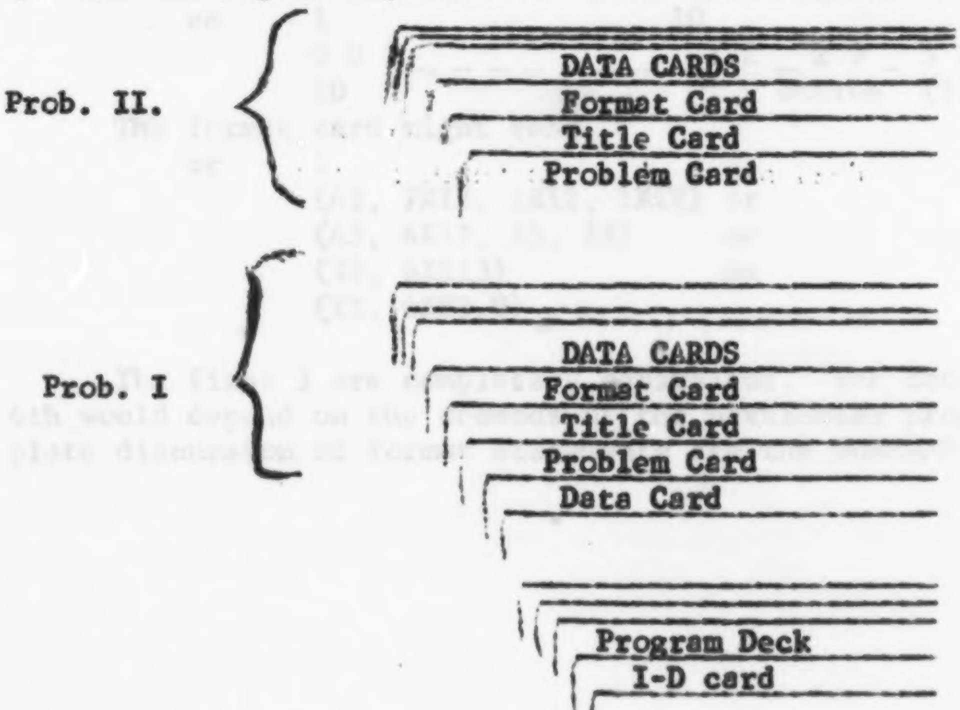
A. Problem Card

B. Title Card

C. Format Card

III. Data Cards

(II and III may be repeated for other sets of data as required.)


Prob. II.
DATA CARDS
Format Card
Title Card
Problem Card
Prob. I
DATA CARDS
Format Card
Title Card
Problem Card
Data Card
Program Deck
I-D card

A Brief Description of Variable Format

An IBM card contains 80 columns (abv. cc). Data is punched into these columns - one number or character per column. If one datum (i.e., score) is represented by two or more digits, two or more columns will be needed. The group of columns representing a datum will be referred to as a field. Before the computer can operate on the data it must be informed of the size (i.e., the number of cc's) and position of these fields in the data cards. Since not all data operated on by the program will be similarly arranged in the data cards, a provision for informing the program as to the data arrangement at object time has been made by utilizing variable format cards. These cards tell the program where the data is in your particular data cards.

The following symbols are used to represent the above information:

(	means the beginning of a format statement; must be present
)	means the end of the format statement, must be present
,	used to separate different types of field designations in the data card
/	means go to new card
A	means the field should be considered a Hollerith character. That is, the field is to be inputted and outputted, but no arithmetic is to be done on the field. This is a way in which letters (A, B, C, etc.) may be manipulated by the computer.
I	represents an integer number. (no decimal) (i.e., 20)
F	represents a fixed point number (may be decimal) (i.e., 20.195)
E	represents a floating point number (exponentially represented) (i.e., .20195-02E = .20195 x 10^{-2})
X	used to indicate a column to be left out

If a data card was punched as follows (_ = blank):

```
cc    1                   10                  20
      0 0 1 _ _ _ _ _ _ _  1 2 _ 2 9 _ 3 2 _ _
      ID                      scores  (3)
```

The format card might read

```
cc    1
      (A3, 7XI2, 1XI2, 1XI2) or
      (A3, 6XI3, I3, I3)     or
      (I3, 6X3I3)            or
      (I3, 6XF3.0)
```

The first 3 are completely equivalent. The choice between the 3rd and 4th would depend on the demands of the particular program. For a more complete discussion of format statements see the FORTRAN manual.

```
      PROGRAM ITEMTS
C*****************************************************************************
C  PROGRAM  ATTANAL   NUC0048         ***** NORTHWESTERN UNIVERSITY
C      DONALD MORRISON   ,   DONALD T. CAMPBELL  ,  LEROY WOLLINS
C     SEPTEMBER 10, 1965
C    COPIES OF THIS PROGRAM MAY BE SECURED FROM THE NORTHWESTERN UNIV. COMPUTING
C CENTER
C*****************************************************************************
      COMMON /DON/ SSD(50,50),SD(50),D(50,50),R(50,50),SX(50),CORSQ(50),
     1E(50),PREV(50)
      COMMON/NAMES/VNAME(50),VAR(50),I,NOSUBJ,TITLE(12),TYPE(50)
      DIMENSION SS(50,50)
      EQUIVALENCE (SS,R)
      REAL INSPB,INDRKR,INRKRA,INRKRB,LA,LAM,L,LX,MAXLAW
      LOGICAL TYPE
      DATA VNMAST,VADD,VTEN,TWENTY,APOS,BPOS,THIRTY,FORTY,FIFTY,GRAND,
     1BLANK/6HITEM00,6H000001,6H000010,6H000020,6HA POS.,6HB POS.,
     26H000030,6H000040,6H000050,6HGRAND ,6H      /
C  ****   FOR CDC 3400 ADD  *******************************************
C      INTEGER VNAME,VNMAST,VADD,VTEN,TWENTY,APOS,BPOS,GRAND,BLANK
C      INTEGER THIRTY,FORTY,FIFTY
C      DATA (VNMAST=8HITEM0000),(VADD=8H00000100),(VTEN=8H00001000)
C      DATA (TWENTY=8H00002000),(APOS=6HA POS.),(BPOS=6HB POS.)
C      DATA (THIRTY=8H00003000),(FORTY=8H00004000),(FIFTY=8H00005000)
C      DATA (GRAND=6HGRAND ),(BLANK=6H      )
 999  FORMAT(1H1,12A6)
 1000 FORMAT(I2,I3,I2,F5.0, 50I1)
 1001 FORMAT(1H ,40X,17HRELIABILITY STUDY,///)
 1002 FORMAT(1H0,20X,23HSPLIT-HALF RELIABILITY=,F6.3)
 1003 FORMAT(1H0,20X,28HKUDER-RICHARDSON FOR TOTAL =,F6.3)
 1004 FORMAT(1H0,20X,34HKUDER-RICHARDSON FOR A POS. TOTAL=,F6.3)
 1005 FORMAT(1H0,20X,34HKUDER-RICHARDSON FOR B POS. TOTAL=,F6.3)
 1006 FORMAT(1H0,38HINFERED AVERAGE INTER-ITEM CORRELATION,//)
 1007 FORMAT(1H0,20X,16HFROM SPLIT-HALF ,2XF6.3)
 1008 FORMAT(1H0,20X,28HFROM KUDER-RICHARDSON(TOTAL),2XF6.3)
 1009 FORMAT(1H0,20X,28HFROM KUDER-RICHARDSON(A POS),2XF6.3)
 1010 FORMAT(1H0,20X,28HFROM KUDER-RICHARDSON(B POS),2XF6.3)
 1011 FORMAT(1H0, 34HCORRECTED ITEM TOTAL CORRELATIONS /)
 1012 FORMAT(33H0ITEM   R(IT)COR.                    )
 1013 FORMAT(1H ,I3, F11.3,9X,A6)
 1014 FORMAT(1H ,20X,34HITERATION TO FIRST FACTOR LOADINGS)
 1015 FORMAT(5H0SUM ,15F7.3)
 1016 FORMAT(5H F    ,15F7.3)
 1017 FORMAT(5H A(I),15F7.3)
 1018 FORMAT(5H0H SQ,15F7.3)
 1019 FORMAT(12A6)
 1020 FORMAT(29H0PREDICTED CORRELATION MATRIX)
 1021 FORMAT(16H0RESIDUAL MATRIX)
 1023 FORMAT(12H A+A+  AVE.=,F6.3)
 1024 FORMAT(12H B+B+  AVE.=,F6.3)
 1025 FORMAT(12H A+B+  AVE.=,F6.3)
 1026 FORMAT(12H TOTAL AVE.=,F6.3)
 1    READ(5,1000)ITEM,NOSUBJ,NOXTR,CONVER,(TYPE(I),I=1,50)
      READ(5,1019)(TITLE(I),I=1,12)
      NOIT=0
      ITOT=ITEM+3
      ITCALL=ITEM+3+NOXTR
      T=ITCALL
      VNAME(1)=VNMAST+VADD
      DO 10 I=2,ITEM
      VNAME(I)=VNAME(I-1)+VADD
      IF(I.EQ.10) VNAME(I)=VNMAST+VTEN
      IF(I.EQ.20)  VNAME(I)=VNMAST+TWENTY
      IF(I.EQ.30) VNAME(I)   =VNMAST+THIRTY
      IF(I.EQ.40) VNAME(I)   =VNMAST+FORTY
      IF(I.EQ.50) VNAME(I)   =VNMAST+FIFTY
10    CONTINUE
      VNAME(ITEM+1)=APOS
      VNAME(ITEM+2)=BPOS
      DO 15 I=ITOT,ITCALL
15    VNAME(I)=BLANK
      VNAME(ITOT)=GRAND
    2 CALL CORREL
      SUBJ1=NOSUBJ-1
      SIGM=0.0
      CONST=(ITEM*(ITEM-1)/2)
      DO 211 I=1,ITEM
      LI=I-1
      IF(LI.LE.0) GO TO 211
      DO 210 J=1,LI
      SIGM=SIGM+(R(I,J)**2)
  210 CONTINUE
  211 CONTINUE
      SIGM=SIGM/CONST
      SIGEX=((1.0-SIGM)**2)/SUBJ1
      SPLTHF=2.*R(ITEM+1,ITEM+2)/(1.+R(ITEM+1,ITEM+2))
      SUMIT =0.0
      SUMITA=0.0
      SUMITB=0.0
      ENA=0.0
      ENB=0.0
      ENI=ITEM
      DO 30 I=1,ITEM
      IF(TYPE(I))GO TO 25
      SUMITB=SUMITB+VAR(I)
      ENB=ENB+1.
      GO TO 30
25    SUMITA=SUMITA+VAR(I)
      ENA=ENA+1.
30    SUMIT=SUMIT+VAR(I)
      RKR=(ENI/(ENI-1.))*(1.-SUMIT/VAR(ITEM+3))
      RKRA=(ENA/(ENA-1.))*(1.-SUMITA/VAR(ITEM+1))
      RKRB=(ENB/(ENB-1.))*(1.-SUMITB/VAR(ITEM+2))
      INSPH=SPLTHF/(ENI+(1.-ENI)*SPLTHF)
      INORKR=RKR/(ENI+(1.-ENI)*RKR)
      INRKRA=RKRA/(ENA+(1.-ENA)*RKRA)
      INRKRB=RKRB/(ENB+(1.-ENB)*RKRB)
C     AVERAGE CORRELATION SECTION
      AASUM=0.0
      BBSUM=0.0
      ABSUM=0.0
      T SUM=0.0
      DO 40 I=1,ITEM
      IP=I+1
      DO 40 J=IP,ITEM
      IF(TYPE(I).AND.TYPE(J)) AASUM=AASUM+R(I,J)
      IF(.NOT.(TYPE(I).OR.TYPE(J))) BBSUM=BBSUM+R(I,J)
      IF((.NOT.TYPE(I).AND.TYPE(J)).OR.(TYPE(I).AND..NOT.TYPE(J)))
     1ABSUM=ABSUM+ R(I,J)
      TSUM=TSUM+R(I,J)
40     CONTINUE
      AASUM=AASUM/(ENA*(ENA-1.)/2.)
      BBSUM=BBSUM/(ENB*(ENB-1.)/2.)
      T SUM=T SUM/(ENI*(ENI-1.)/2.)
      ABSUM=ABSUM/((ENI*(ENI-1.)-ENA*(ENA-1.)-ENB*(ENB-1.))/2.)
      WRITE(6,999)(TITLE(I),I=1,12)
      WRITE(6,1001)
      WRITE(6,1002)SPLTHF
      WRITE(6,1003)RKR
      WRITE(6,1004)RKRA
      WRITE(6,1005)RKRB
      WRITE(6,1006)
      WRITE(6,1007)INSPH
      WRITE(6,1008)INDRKR
      WRITE(6,1009)INRKRA
      WRITE(6,1010)INRKRB
      WRITE(6,1023)AASUM
      WRITE(6,1024)BBSUM
      WRITE(6,1025)ABSUM
      WRITE(6,1026)T SUM
      WRITE(6,1011)
      WRITE(6,1012)
      DO 50 I=1,ITEM
      COREC=(R(I,ITOT)*SD(ITOT)-SD(I))
      COREC=COREC/(SQRT(VAR(ITOT)    +VAR(I)    -2.*R(ITOT,I)*SD(ITOT
     1*SD(I)))
      CORSQ(I)=COREC**2
      IF(TYPE(I)) MODE=APOS
      IF(.NOT.TYPE(I)) MODE=BPOS
50    WRITE(6,1013)I,COREC,        MODE
C     FACTOR ANALYSIS SECTION
C     COMPUTATION OF FIRST FACTOR LOADINGS
      WRITE(6,999)(TITLE(I),I=1,12)
      WRITE(6,1014)
      DO 60 I=1,ITEM
      PREV(I)=CORSQ(I)
60    SA(I)=0.0
      DO 70 I=1,ITEM
      DO 70 J=1,ITEM
      IF(I.EQ.J) GO TO 70
      SA(I)=SA(I)+R(I,J)
70    CONTINUE
      WRITE(6,1015)(SA(I),I=1,ITEM)
      DO 75 I=1,ITEM
      IF(SA(I).LT.0.0)     SA(I)=0.0
75    CONTINUE
80    SUM=0.0
      WRITE(6,1018)(PREV(I),I=1,ITEM)
      NOIT=NOIT+1
      DO 90 I=1,ITEM
      E(I)=SA(I)*PREV(I)
      SUM=SUM+E(I)
90    CONTINUE
      WRITE(6,1016)(E(I),I=1,ITEM)
      ENDSW=0.0
      SUM=1.0/SQRT(SUM)
      DO 100 I=1,ITEM
      E(I)=E(I)*SUM
      IF((ABS(E(I)*E(I)-PREV(I))).GT.(.01 ))ENDSW=1.0
      PREV(I)=E(I)*E(I)
100   CONTINUE
      WRITE(6,1017)(E(I),I=1,ITEM)
      IF(NOIT.GT.19) GO TO 99
      IF(ENDSW-.5)99,99,80
99    DO 120 I=1,ITEM
      DO 120 J=1,ITEM
      D(I,J)=E(I)*E(J)
      SSD(I,J)=R(I,J)-D(I,J)
      IF(I.EQ.J)  D(I,J)=0.0
      IF(I.EQ.J)SSD(I,J)=0.0
120    CONTINUE
      WRITE(6,1018)(PREV(I),I=1,ITEM)
      WRITE(6,999)(TITLE(I),I=1,12)
      WRITE(6,1020)
      CALL MPRINT(D,ITEM,1,6HPRED R)
      WRITE(6,1021)
      CALL MPRINT(SSD,ITEM,1,5HRESID)
      SIG1=0.0
      SIGO=0.0
      DO 208 I=1,ITEM
      LI=I-1
      IF(LI.LE.0) GO TO 208
      DO 200 J=1,LI
      SIGO=SIGO+(SSD(I,J)**2)
      SIG1=SIG1 + SSD(I,J)
  200 CONTINUE
  208 CONTINUE
      SIGO=SQRT(SIGO/CONST)
      SIG1=SIG1/CONST
      WRITE (6,215) SIG1,SIGO
  215 FORMAT(1H1,30X,21H MEAN RESIDUAL EQUALS F10.4,21HSIGMA OF RESIDUAL
     1S IS  F10.4)
      CALL APTERX(NOSUBJ,SIGM)
      GO TO 1
       END
      SUBROUTINE APTEHX(NOSUBJ,SIGM)
C *****  FOR CDC 3600 ADD NOTHING TO THIS SUBROUTINE  ******************
      DIMENSION A(50,50), F(50,50), DV(50), C(50), W(50), W1(50), R(50),
     1 MORA(50), MCOM(50), MORW1(50), MORW2(50),DUMMY(2300)
      COMMON /DON/ A,F,DV,C,W,W1,R,MORA,MCOM,MORW1,MORW2,DUMMY           IOWA0120
  902 WRITE (6,99)
   99 FORMAT(1H1/46HNORTHWESTERN UNIVERSITY MAXIMUM LIKLIHOOD PROGRAM
     1TO EXTRACT ONE FACTOR)
  100 CONTINUE
      CONVDR=.01
      I=IPR2
      CONST=0.01                                                          IOWA0170
      KUIT=0                                                              IOWA0190
      SS=1E9                                                              IOWA0200
      RNIT=0
      M=1
      REWIND 3
C     READ TAPE 3 FOR CORRELATION MATRIX
   17 READ(3)N,((A(I,J),J=1,N),I=1,N)
      REWIND 3                                                            IOWA0440
   98 DO 12 I=1,N
   12 C(I)=.0
C *****     PRINCIPAL COMPONENTS SECTION                  ****************
  820 DO 822 I=1,N
  822 DV(I)=A(I,I)
      DO 814 I=1,N
      W(I)=DV(I)-C(I)
  814 W1(I)=DV(I)-C(I)
      KL=N-1
      DO 821 J=1,KL
      L=J+1
      DO 821 IK=L,N
  821 A(IK,J)=A(J,IK)
      CALL HOW(N,50,1,A,W,F,MORA,MCOM,MORW1,MORW2)
      DO 833 I=1,N
      C(I)=0.0
      DO 833 J=1,M
  833 C(I)=C(I)+(F(I,J)*F(I,J))*W(J)
      WRITE(6,840)(C(I),I=1,N)
  840 FORMAT(1H0,40HPRINCIPAL COMPONENTS COMMUNALITIES  (8F10.4))
      READ (3)N,((A(I,J),J=1,N),I=1,N)
      REWIND 3
      NCAN2=20                                                            IOWA0520
   20 DO 22 I=1,N                                                         IOWA0530
   22 DV(I)=A(I,I)                                                        IOWA0540
      DO 14 I=1,N                                                         IOWA0550
      W(I)=DV(I)-C(I)                                                     IOWA0560
   14 W1(I)=W(I)                                                          IOWA0570
  200 NUIT=NUIT+1                                                         IOWA0730
   24 DO 27 I=1,N                                                         IOWA0740
   27 A(I,I)=DV(I)-W(I)                                                   IOWA0760
   29 DO 30 I=1,N                                                         IOWA0770
   30 A(I,I)=A(I,I)/W(I)                                                  IOWA0780
      DO 31 J=2,N                                                         IOWA0790
      KJ=J-1                                                              IOWA0800
      DO 31 I=1,KJ                                                        IOWA0810
   31 A(J,I)=A(I,J)/SQRT(W(I)*W(J))
```

```
                                                                        I
   14 CALL [illegible] (N, NO, M, A, H, F, HOWA, HOWB, HOWW1, HOWW2)    IOWA0830
      DO 43 I=1,N                                                       IOWA0840
      C(I)=0.                                                           IOWA0850
      DO 43 J=1,M                                                      .IOWA0860
   43 C(I)=C(I)+(F(I,J)*F(I,J))*H(J)                                    IOWA0880
   46 IF(I[illegible]-1) 48,37,800                                      IOWA0890
  800 DO 801 I = 1,N                                                    IOWA0900
      W(I)   = UV(I) / (UV(I) + C(I))                                   IOWA0910
  801 C(I)   = [illegible](I) - W(I)                                    IOWA0920
      GO TO 98                                                          IOWA0930
   37 FL=0.                                                             IOWA0940
      DO 700 I=1,N                                                      IOWA0950
  700 FL=FL+1.0/W(I)                                                    IOWA0960
      DO 701 I=1,N                                                      IOWA0970
  701 [illegible]=W(I)                                                  IOWA0980
      A =                                                               IOWA0990
      B =                                                               IOWA1000
      F[illegible]/(FN[illegible])                                      IOWA1010
      DO 702 I=1,N                                                      IOWA1020
      C(I)=C(I)*W(I)                                                    IOWA1030
      W(I)=0.                                                           IOWA1040
      DO 703 J=1,M                                                      IOWA1050
  703 W(I)=W(I)+(F(I,J)*F(I,J))*((H(J)+1.0-FL)/FL)                      IOWA1060
  702 W(I)=1.0/(1.0+W(I))                                               IOWA1070
      GO TO 98                                                          IOWA1080
   98 DO 42 I=1,N                                                       IOWA1090
      C(I)=C(I)*W(I)                                                    IOWA1100
   41 W(I)=UV(I)-C(I)
   40 CONTINUE
   42 IF(NOIT.LT.MAXW[illegible]) GO TO 44                              IOWA1180
   45 NOIT=1                                                            IOWA1190
      GO TO 900                                                         IOWA1250
   44 L=0
C ********* CHECK FOR NEGATIVE LOADINGS ***********************************
      DO 658 J=1,N
      IF(F(J,1).LT.0.0) F(J,1)=0.0
  658 CONTINUE                                                          IOWA1260
      DO 50 I=1,N                                                       IOWA1270
      IF(ABS(W(I)-W1(I))-CONV)50,50,51                                  IOWA1280
   51 L=1                                                               IOWA1290
   50 CONTINUE                                                          IOWA1300
      IF(L[illegible]2,45,52                                            IOWA1310
   52 DO 53 I=1,N                                                       IOWA1320
   53 W1(I)=W(I)
      GO TO [illegible]
  300 CONTINUE                                                          IOWA1400
   55 DO 56 I=1,M                                                       IOWA1410
      DO 56 J=1,N                                                       IOWA1420
   56 F(J,I)=F(J,I)*SQRT(W(J)*H(I))                                     IOWA1430
   57 SS=0.
      SN=0.0
      [illegible]=0.0                                                   IOWA1440
      S1=0.                                                             IOWA1450
      DO 59 I=1,N1                                                      IOWA1460
      I1=I+1                                                            IOWA1470
      DO 59 J=I1,N                                                      IOWA1480
      [illegible]
      DO 60 K=1,M                                                       IOWA1490
   60 TEMP=TEMP+F(I,K)*F(J,K)                                           IOWA1500
      A(J,I)=A(I,J)-TEMP                                                IOWA1510
      DEP=DEP+(1.0/(W(I)*W(J)))
      SS =SSM+(A(J,I)*A(J,I))
   52 SS =SSM+(A(J,I)*A(J,I))/( W(I)*W(J)*W(I)*W(J))
   59 CONTINUE                                                          IOWA1560
      SSH=2.0*SSH                                                       IOWA1570
   70 WRITE (6,80)NOIT,SSH
   80 FORMAT(1H1,    2X,22HFINAL RESULTS --          5X,21HNUMBER OF ITE
     1RATIONS=,I5,5X,24HRESIDUAL SUM OF SQUARES=,E16.8)                 IOWA1810
   75 DO 78 I=1,N                                                       IOWA1860
   78 A(I,I)=DV(I)                                                      IOWA1870
      WRITE (6,68)                                                      IOWA1880
   68 FORMAT(1H0,13HCOMMUNALITIES)                                      IOWA1680
      WRITE (6,69)(C(I),I=1,N)                                          IOWA1890
   69 FORMAT(1H0 8F16.6)
      WRITE (6,79)                                                      IOWA1900
   79 FORMAT(1H0,18HSPECIFIC VARIANCES  )                               IOWA1910
      WRITE (6,69)(W(I),I=1,N)                                          IOWA1920
   77 WRITE (6,81)                                                      IOWA1940
   81 FORMAT(1H0,15HFACTOR LOADINGS)                                    IOWA1950
      DO 82 I=1,M                                                       IOWA1960
   82 WRITE (6,69)(F(J,I),J=1,N)                                        IOWA1970
      WRITE (6,83)                                                      IOWA1980
   83 FORMAT(1H0,55H           CORRELATION INPUT MATRIX AND RESIDUAL MATIOWA1990
     1RIX)                                                              IOWA2000
      DO 97 I=1,N                                                       IOWA2010
   97 WRITE (6,69)(A(I,J),J=1,N)                                        IOWA2020
C *********  LAWLEY TEST *************************************************
      OSUBJ=NOSUBJ
      ENI=N
      CHILAW=(SSH/2.0)*(OSUBJ-.166*(2.0*ENI+5.0)-1.66)
      NDF=(N**2-3*N)/2
      WRITE(6,1022)CHILAW,NDF
 1022 FORMAT(28H CHI-SQUARE FOR LAWLEY TEST=,F10.5,20HNO. DEG. FREEDOM I
     1S   I4)
      CALL CHISQ(VAL,NDF)
      ABARSQ=0.0
      DO 750 I=1,N
  750 ABARSQ=ABARSQ+(1.0/(W(I)**2))
      ABARSQ=ABARSQ/ENI
      DENOM=(ENI*ABARSQ*(ENI*ABARSQ-1.0)/2.0)
      ENUM1=VAL/(OSUBJ-.33*ENI-2.5)
      SSH=SSH/2.0
      CEF=SSH-((DEP*SSM)/(.5*ENI*(ENI-1.0)))
      ENUM=ENUM1-CEF
      WOLEA=ENUM/DENOM
      SSM=SSM/(ENI*(ENI-1.0))
      SIGM=SQRT(SIGM)
      SSM=SQRT(SSM)
      WOLEA=SQRT(WOLEA)
      WOLIN=(SIGM- SSM)/(SIGM-WOLEA)
      WRITE(6,215)SIGM, SSM,WOLEA,WOLIN
  215 FORMAT(1H1,40X,54H                    WOLINS  INDEX OF SINGLE FACTORED
     1DNESS  /    26H0 SIGMA    =              F10.4,28H   SIGMA OF RESI
     1DUALS EQUALSF10.4,28H * EXPECTED  SIGMA          F10.4//16H   INDE
     1X EQUALS    F10.4)
   86 RETURN
      END
      SUBROUTINE CORREL
C     ******   FOR CDC 3400 ADD   ********************************************
      INTEGER VNAME
      COMMON /DON/ SSD(50,50),SD(50),D(50,50),R(50,50),SX(50),CORSQ(50),
     1E(50),PREV(50)
      COMMON/NAMES/VNAME(50),VAR(50),T,NOSUBJ,TITLE(12),TYPE(50)
      DIMENSION SS(50,50),FMT(12)
      DIMENSION X(50),XM(50)
      EQUIVALENCE (SS,R)
      LOGICAL TYPE
 1000 READ(5,1)(FMT(I),I=1,12)
    1 FORMAT(12A6)
      M=T                                                               0610
  210 ENG=NOSUBJ
      DO 230  I=1,M                                                     0640
  220 SX(I)=0.0                                                         0650
      DO 230  J=1,M                                                     0660
  230 SS(I,J)=0.0                                                       0670
      CASES=ENG                                                         0680
  240 READ(5,FMT)(X(J),J=1,M )
      DO 260 I = 1,M                                                    0700
      SX(I)=SX(I)+X(I)                                                  0710
      DO 260 J = I,M                                                    0720
  260 SS (I,J) = SS(I,J) + X (I)* X(J)                                  0730
C     RAW SUM OF SQUARES AND CROSS-PRODUCTS                             0740
      CASES=CASES-1.0                                                   0750
      IF (CASES)  280,280,240                                           0760
  280 DO 286  I=1,M                                                     0770
      DO 286  J=I,M                                                     0780
      SSD(I,J)=SS(I,J)-SX(I)*SX(J)/ENG                                  0790
      SS(J,I) = SS(I,J)                                                 0800
  286 SSD(J,I)= SSD(I,J)                                                0810
C     DEVIATION SUM OF SQUARES AND CROSS-PRODUCTS                       0820
      DO 295  I=1,M                                                     0830
      XM(I)=SX(I)/ENG                                                   0840
  295 SD(I)=SQRT(SSD(I,I)/(ENG-1.0))                                    0850
C     MEANS AND STANDARD DEVIATIONS                                     0860
      WRITE (6,31)(TITLE(I), I=1,9 )
 31   FORMAT(1H1,8A6,A5)
      WRITE (6,32)T                                                     0890
   32 FORMAT (23H0NUMBER OF VARIABLES = , F3.0)                         0900
      NG=ENG
      WRITE (6,33)NG                                                    0910
   33 FORMAT (36H0NUMBER OF SUBJECTS IN THIS GROUP = , I5)              0920
      WRITE(6,499)
 499  FORMAT(60H VARIABLE   MEAN   STD. DEV. VARIANCE     SUM    SUM OF
     1SQS.)
      DO 500 I=1,M
      VAR(I)=SD(I)*SD(I)
 500  WRITE(6,501) VNAME(I),XM(I),SD(I),VAR(I),SX(I),SS(I,I)
 501  FORMAT(1H ,1X,A6,1X,3(2XF7.3),2(F9.2,1X))
  435 DO 441  I=1,M                                                     1240
      DO 441  J=I,M                                                     1250
      D(I,J)=SSD(I,J)/(ENG-1.0)                                         1260
  441 D(J,I)=D(I,J)                                                     1270
C     DISPERSION MATRIX                                                 1280
  475 DO 486 I = 1,M                                                    1370
```

```
      DO 486  J=1,M
      R(I,J) = D (I,J)/(SD(I)* SD(J))
  486 R(J,I)=R(I,J)
C     CORRELATION MATRIX
  625 IPNO=1
  628 WRITE (6,50)
   50 FORMAT (20H CORRELATION MATRIX,)
      CALL MPRINT(R,M,1,5HR MAT)
      M=M
      WRITE(3)M,((R(I,J),J=1,M),I=1,M)
      REWIND 3
  700 RETURN
       END
      SUBROUTINE EIGVA(LP,NM,M,E,A,B,F,W)                               EI0020
C *** FOR CDC 3400 ADD NOTHING TO THIS SUBROUTINE ******************
C EIGEN VALUE SUBROUTINE FOR TRI-DIAGONAL MATRICES     DWM     1517-08  EI0010
      DIMENSION E(50), A(50), B(50), F(50), W(50)
      EQUIVALENCE (S1,IS1),(S2,IS2)                                     EI0050
C  FIND UPPER AND LOWER BOUNDS AND NORMALIZE INPUT                      EI0060
      BD=ABS(A(1))                                                      EI0070
      DO 5 I=2,LP                                                       EI0080
    5 BD=AMAX1(BD,ABS(A(I))+B(I)**2)                                    EI0090
      BD=BD+1.0                                                         EI0100
      DO 6 I=1,LP                                                       EI0110
      A(I)=A(I)/BD                                                      EI0120
      B(I)=B(I)/BD                                                      EI0130
      W(I)=1.0                                                          EI0140
    6 E(I)=-1.0                                                         EI0150
      DO 50 K=1,LP                                                      EI0160
    8 IF((W(K)-E(K))/AMAX1(ABS(W(K)),ABS(E(K)),1.0E-9)-1.0E-7)50,50,10  EI0170
   10 X=(W(K)+E(K))*.5                                                  EI0180
C  FIND NUMBER OF EIGEN VALUES ,N, GREATER THAN OR EQUAL TO X           EI0190
      IS2=1                                                             EI0200
      F(1)=A(1)-X                                                       EI0210
      IF(F(1))102,104,104                                               EI0220
  102 IS1=-1                                                            EI0230
      N=0                                                               EI0240
      GO TO 105                                                         EI0250
  104 IS1=1                                                             EI0260
      N=1                                                               EI0270
  105 DO 120 I=2,LP                                                     EI0280
      IF(B(I))106,113,106                                               EI0290
  106 IF(B(I-1))107,114,107                                             EI0300
  107 IF(ABS(F(I-1))+ABS(F(I-2))-1.0E-15)110,112,112                    EI0310
  110 F(I-1)=F(I-1)*1.0E15                                              EI0320
      F(I-2)=F(I-2)*1.0E15                                              EI0330
  112 F(I)=(A(I)-X)*F(I-1)-B(I)**2*F(I-2)                               EI0340
      GO TO 115                                                         EI0350
  113 F(I)=(A(I)-X)*SIGN(1.0,S1)                                        EI0360
      GO TO 115                                                         EI0370
  114 F(I)=(A(I)-X)*F(I-1)-SIGN(B(I)**2,S2)                             EI0380
  115 S2=S1                                                             EI0390
      IF(F(I))116,117,116                                               EI0400
  116 S1=SIGN(S1,F(I))                                                  EI0410
      IF(IS2*IS1)117,120,117                                            EI0420
  117 N=N+1                                                             EI0430
  120 CONTINUE                                                          EI0440
C  TRAP EIGEN VALUES IN SMALLER AND SMALLER BOUNDS                      EI0450
      N=LP-N                                                            EI0460
      IF(N-K)20,12,12                                                   EI0470
   12 DO 15 J=K,N                                                       EI0480
   15 W(J)=X                                                            EI0490
   20 N=N+1                                                             EI0500
      IF(LP-N)8,24,24                                                   EI0510
   24 DO 26 J=N,LP                                                      EI0520
      IF(X-E(J))8,8,26                                                  EI0530
   26 E(J)=X                                                            EI0540
      GO TO 8                                                           EI0550
   50 CONTINUE                                                          EI0560
C  RESTORE INPUT AND ORDER EIGEN VALUES                                 EI0570
```

```
      DO 60 I=1,LP                                                       EI0580
      A(I)=A(I)*HD                                                       EI0590
      B(I)=B(I)*HD                                                       EI0600
   60 F(I)=(W(I)+E(I))*HD*.5                                             EI0610
      J=LP                                                               EI0620
      K=1                                                                EI0630
      DO 70 I=1,LP                                                       EI0640
      IF(F(K) - F(J)) 63,63,65                                           EI0650
   63 E(I)=F(J)                                                          EI0660
      J=J-1                                                              EI0670
      GO TO 70                                                           EI0680
   65 E(I)=F(K)                                                          EI0690
      K=K+1                                                              EI0700
   70 CONTINUE                                                           EI0710
      IF(ISIGN(1,M))75,80,80                                             EI0720
   75 DO77 I=1,LP                                                        EI0730
   77 F(I)=E(I)                                                          EI0740
      J=LP                                                               EI0750
      DO 79 I=1,LP                                                       EI0760
      E(I)=F(J)                                                          EI0770
   79 J=J-1                                                              EI0780
   80 CONTINUE                                                           EI0790
      RETURN                                                             EI0800
       END
      SUBROUTINE EIGVE(LP,NM,R,A,B,E,V,P,Q)
C ****  FOR CDC 3400 ADD NOTHING TO THIS SUBROUTINE  *******************
      DIMENSION R(2500), A(50), B(50), V(2500), P(50), Q(50)
C  SET UP SIMULTANEOUS EQUATIONS FOR EIGEN VECTOR WITH EIGEN VALUE E
      X=A(1)-E
      Y=B(2)
      LP1=LP-1
      DO 10 I=1,LP1
      IF(ABS(X)-ABS(B(I+1)))4,6,6
    4 P(I)=B(I+1)
      Q(I)=A(I+1)-E
      V(I)=B(I+2)
      Z=-X/P(I)
      X=Z*Q(I)+Y
      IF(LP1-I)5,10,5
    5 Y=Z*V(I)
      GO TO 10
    6 IF(X)8,7,8
    7 X=1.0E-10
    8 P(I)=X
      Q(I)=Y
      V(I)=0.0
      X=A(I+1)-(B(I+1)/X*Y+E)
      Y=B(I+2)
   10 CONTINUE
C  SOLVE SIMULTANEOUS EQUATIONS FOR EIGEN VECTOR OF TRI-DIAGONAL MATRIX
   20 IF(X) 21,28,21
   21 V(LP)=1.0/X
   22 I=LP1
      V(I)=(1.0-Q(I)*V(LP))/P(I)
      X=V(LP)**2+V(I)**2
   25 I=I-1
      IF(I) 26,30,26
   26 V(I)=(1.0-(Q(I)*V(I+1)+V(I)*V(I+2)))/P(I)
      X=X+V(I)**2
      GO TO 25
   28 V(LP)=1.0E10
      GO TO 22
   30 X=SQRT(X)
      DO 31 I=1,LP
   31 V(I)=V(I)/X
C  TRANSFORM EIGEN VECTOR TO SOLUTION OF ORIGINAL MATRIX
      J=LP1*NM-NM
      K=LP
      GO TO 42
   32 K=K-1
      J=J-NM
      Y=0.0
      DO 35 I=K,LP
      L=J+I
   35 Y=Y+V(I)*R(L)
      DO 40 I=K,LP
      L=J+I
   40 V(I)=V(I)-Y*R(L)
   42 IF(J)32,44,32
   44 RETURN
       END
      SUBROUTINE HOW(LP,NM,M,R,E,V,A,B,W1,W2)
C**** FOR CDC 3400 ADD NOTHING TO THIS SUBROUTINE ******************
      DIMENSION R(2500), E(50),V(2500), A(50), B(50), W1(50), W2(50)
      IF(LP-1)20,14,4
C  TRI-DIAGONALIZE MATRIX
    4 CALL TRIDI(LP,NM,R(1),A(1),B(1),W1(1),W2(1))
C  FIND EIGEN VALUES
      CALL EIGVA(LP,NM,M,E(1),A(1),B(1),W1(1),W2(1))
      IF(M)5,9,5
C  FIND EIGEN VECTORS
    5 K=IABS(M)
      J=1
      DO 7 I=1,K
      CALL EIGVE(LP,NM,R(1),A(1),B(1),E(I),V(J),W1(1),W2(1))
    7 J=J+NM
C  RESTORE INPUT MATRIX
    9 NM1=NM+1
      JJ=NM1
      LP2=LP*NM
      DO 12 I=2,LP2,NM1
      K=I
      DO 10 J=JJ,LP2,NM
      R(K)=R(J)
   10 K=K+1
   12 JJ=JJ+NM1
      GO TO 20
C  SOLUTION FOR ORDER 1
   14 E(1)=R(1)
      V(1)=1.0
      A(1)=R(1)
      B(1)=0.0
   20 RETURN
       END
      SUBROUTINE  MPRINT (R, M, L, TI)
****** FOR CDC 3400 AUD  **************************************
      INTEGER VNAME
C     SUBROUTINE MPRINT
C          THIS SUBROUTINE WILL PRINTOUT SQUARE MATRICES UP TO 50 X 50.
C     ALL ROWS AND COLUMNS ARE APPROPRIATELY LABELED.
C          DEFINITITIONS OF SUBROUTINE ARGUMENTS.
C     R(I,J) = MATRIX TO BE PRINTED
C     M = ORDER (MAX. OF 50)
C     L = 1 FOR 15F7.3 OUTPUT OR 2 FOR 10F11.3
C     TI IS A SIX CHARACTER DESIGNATE OF THE MATRIX
      COMMON/NAMES/VNAME(50),VAR(50),T,NOSUBJ,TITLE(12),TYPE(50)
      DIMENSION  R(50,50)
      LOGICAL TYPE
      IF(L-2)2,4,2
    2 LI=14
      N = 15
      GO TO 5
    4 LI=9
      N = 10
    5 J1=0
      J2=0
      JSEC = 0
    9 J1 = J2 +1
      J2 = J1 +LI
      IF (J2 - M) 13,13,12
   12 J2=M
   13 JSEC = JSEC + 1
      IF (JSEC -1)  18, 18 , 19
   18 WRITE (6,17)TI, JSEC
   17 FORMAT (1H0, A6, 9H  SECTION I3/)
      GO TO  201
   19 WRITE (6,20)TI,JSEC
   20 FORMAT (1H1, A6, 9H  SECTION I3/)
  201 IF(L-2)21,26,99
   21 WRITE (6,22)(VNAME(I),I=J1,J2)
   22 FORMAT(6H0  VAR 9X,15(A6,1X))
   99 DO 23 I=1,M
   23 WRITE (6,24) VNAME(I),(R(I,K),K=J1,J2)
   24 FORMAT(1H A6,7X,15F7.3)
      GO TO 31
   26 WRITE (6,27)(VNAME(I),I=J1,J2)
   27 FORMAT(6H0  VAR 9X,10(A6,5X))
      DO 29 I=1,M
   29 WRITE (6,30) VNAME(I),(R(I,K),K=J1,J2)
   30 FORMAT (1H A6,4X,10F11.2)
   31 IF (J2-M) 9,32,32
   32 RETURN
       END
      SUBROUTINE TRIDI(LP,NM,H,A,B,W,Q)                                    TR0060
C **** FOR CDC 3400 ADD NOTHING TO THIS SUBROUTINE  *******************
C  TRI-DIAGONALIZATION SUBROUTINE       DWM        1517-08                 TR0010
C  RETURN ORIGNAL H IN UPPER TRIANGULAR HALF INCLUDING DIAGONAL           TR0020
C  RETURN MODIFIED W MATRICES IN LOWER HALF OF H MATRIX                   TR0030
C  RETURN NEW DIAGONAL IN A                                               TR0040
C  RETURN NEW FIRST OFF DIAGONAL IN B                                     TR0050
      DIMENSION H(2500), A(50), B(50), Q(50), W(50)
      LP1=LP-1                                                             TR0090
      LP2=LP1*NM+LP                                                        TR0100
      LPP=LP2-NM                                                           TR0110
      NM1=NM+1                                                             TR0120
C  STORE ORIGINAL DIAGONAL                                                TR0130
      L=0                                                                  TR0140
      DO 10 I=1,LP2,NM1                                                    TR0150
      L=L+1                                                                TR0160
   10 A(L)=H(I)                                                            TR0170
      B(1)=0.0                                                             TR0180
      IF(LP-2)99,65,15                                                     TR0190
   15 KK=0                                                                 TR0200
      DO 50 K=2,LP1                                                        TR0210
      KL=KK+K                                                              TR0220
      KU=KK+LP                                                             TR0230
      KJ=K+1                                                               TR0240
C  CALCULATE AND STORE MODIFIED COLUMN MATRIX W                           TR0250
      SUM=0.0                                                              TR0260
      DO 20 J=KL,KU                                                        TR0270
   20 SUM=SUM+H(J)**2                                                      TR0280
      S=SQRT(SUM)                                                          TR0290
      B(K)=SIGN(S,-H(KL))                                                  TR0300
      S=1.0/S                                                              TR0310
      W(K)=SQRT(ABS(H(KL))*S+1.0)                                          TR0320
      X=SIGN(S/W(K),H(KL))                                                 TR0330
      H(KL)=W(K)                                                           TR0340
      DO 30 I=KJ,LP                                                        TR0350
      JJ=I+KK                                                              TR0360
      W(I)=X*H(JJ)                                                         TR0370
   30 H(JJ)=W(I)                                                           TR0380
C  CALCULATE NEW H MATRIX WITH ROW K-1 NOW HAVING ZEROS OFF 2ND DIAGONAL  TR0390
      DO 35 J=K,LP                                                         TR0400
      JJ=J+1                                                               TR0410
      Q(J)=0.0                                                             TR0420
      L=KK+J                                                               TR0430
      DO 33 I=K,J                                                          TR0440
      L=L+NM                                                               TR0450
   33 Q(J)=Q(J)+H(L)*W(I)                                                  TR0460
      IF(JJ-LP)34,34,36                                                    TR0470
   34 DO 35 I=JJ,LP                                                        TR0480
      L=L+1                                                                TR0490
   35 Q(J)=Q(J)+H(L)*W(I)                                                  TR0500
   36 X=0.0                                                                TR0510
      DO 40 J=K,LP                                                         TR0520
   40 X=X+W(J)*Q(J)                                                        TR0530
      X=.5*X                                                               TR0540
      DO 45 I=K,LP                                                         TR0550
   45 Q(I)=X*W(I)-Q(I)                                                     TR0560
      LL=KK                                                                TR0570
      KK=KK+NM
      DO 50 I=K,LP
      LL=LL+NM
      DO 50 J=I,LP
      L=LL+J
   50 H(L)=H(L)+Q(I)*W(J)+Q(J)*W(I)
C     SORT OUTPUT
      L=1
      DO 60 I=1,LP
      X=A(I)
      A(I)=H(L)
      H(L)=X
   60 L=L+NM1
   65 B(LP)=H(LPP)
   99 RETURN
       END
      SUBROUTINE CHISQ (VAL,N)
C *****  FOR CDC 3400 ADD NOTHING TO THIS SUBROUTINE  *******************
      DIMENSION C(30)
      C(1)=.455
      C(2)=1.386
      C(3)=2.366
      C(4)=3.357
      C(5)=4.351
      C(6)=5.348
      C(7)=6.346
      C(8)=7.344
      C(9)=8.343
      C(10)=9.342
      C(11)=10.341
      C(12)=11.340
      C(13)=12.340
      C(14)=13.339
      C(15)=14.339
      C(16)=15.338
      C(17)=16.338
      C(18)=17.338
      C(19)=18.338
      C(20)=19.337
      C(21)=20.337
      C(22)=21.337
      C(23)=22.337
      C(24)=23.337
      C(25)=24.337
      C(26)=25.336
      C(27)=26.336
      C(28)=27.336
      C(29)=28.336
      C(30)=29.336
      VAL=C(N)
      RETURN
      END
```

#5

```
077201    22  ENGINEERING MANAGEMENT    J. UTTERBACK
NUMBER OF VARIABLES =  11

NUMBER OF SUBJECTS IN THIS GROUP =     72
VARIABLE    MEAN    STD. DEV. VARIANCE      SUM      SUM OF SQS.
 ITEM01      1.903     1.416     2.004    137.00     403.00
 ITEM02      2.556     1.255     1.574    184.00     582.00
 ITEM03      2.333     1.538     2.366    168.00     560.00
 ITEM04      2.569     1.161     1.347    185.00     571.00
 ITEM05      2.278     1.567     2.457    164.00     548.00
 ITEM06      2.153     1.526     2.328    155.00     499.00
 ITEM07      1.875     1.186     1.407    135.00     353.00
 ITEM08      2.333     1.404     1.972    168.00     532.00
 A POS.      9.444     3.845    14.786    680.00    7472.00
 B POS.      8.556     3.623    13.124    616.00    6202.00
 GRAND      18.242     6.685    44.689   1313.40   27131.94
CORRELATION MATRIX.

R MAT   SECTION  1

  VAR          ITEM01 ITEM02 ITEM03 ITEM04 ITEM05 ITEM06 ITEM07 ITEM08 A POS. B POS. GRAND
ITEM01          1.000  0.174  0.474  0.197  0.526  0.143  0.118  0.544  0.507  0.733  0.692
ITEM02          0.174  1.000  0.244  0.089  0.244  0.190  0.274  0.245  0.547  0.325  0.500
ITEM03          0.474  0.244  1.000  0.105  0.514  0.224  0.062  0.613  0.765  0.492  0.710
ITEM04          0.197  0.089  0.105  1.000  0.020  0.099  0.216  0.202  0.195  0.453  0.380
ITEM05          0.526  0.244  0.514  0.020  1.000  0.212  0.307  0.404  0.570  0.776  0.747
ITEM06          0.143  0.190  0.224  0.099  0.212  1.000  0.143  0.295  0.620  0.290  0.498
ITEM07          0.118  0.274  0.062  0.216  0.307  0.143  1.000  0.195  0.229  0.590  0.447
ITEM08          0.544  0.245  0.613  0.202  0.404  0.295  0.195  1.000  0.765  0.636  0.784
A POS.          0.507  0.547  0.765  0.195  0.570  0.620  0.229  0.765  1.000  0.800  0.900
B POS.          0.733  0.325  0.492  0.453  0.776  0.290  0.590  0.636  0.800  1.000  0.887
GRAND           0.692  0.500  0.710  0.380  0.747  0.498  0.447  0.784  0.900  0.887  1.000
```

46

```
0807201   22  ENGINEERING MANAGEMENT   J. UITERHACK          01100101
                                         RELIABILITY STUDY


                          SPLIT-HALF RELIABILITY= 0.750

                          KUDER-RICHARDSON FOR TOTAL = 0.748

                          KUDER-RICHARDSON FOR A POS. TOTAL= 0.726

                          KUDER-RICHARDSON FOR B POS. TOTAL= 0.447

INFERED AVERAGE INTER-ITEM CORRELATION


                          FROM SPLIT-HALF     0.273

                          FROM KUDER-RICHARDSON(TOTAL)    0.270

                          FROM KUDER-RICHARDSON(A POS)    0.399

                          FROM KUDER-RICHARDSON(B POS)    0.168
A*A*   AVE.= 0.246
B*B*   AVE.= 0.275
A*B*   AVE.= 0.260
TOTAL  AVE.= 0.250

CORRECTED ITEM TOTAL CORRELATIONS


ITEM    R(IT)COR.
  1       0.554         A POS.
  2       0.340         B POS.
  3       0.563         B POS.
  4       0.196         A POS.
  5       0.610         B POS.
  6       0.295         B POS.
  7       0.249         A POS.
  8       0.679         A POS.
```

```
0807201   22  ENGINEERING MANAGEMENT    J. UTTERBACK         01100101
                          ITERATION TO FIRST FACTOR LOADINGS

SUM     2.221   1.483   2.240   0.899   2.449   1.287   1.315   2.671

H SQ    0.307   0.115   0.317   0.038   0.372   0.087   0.083   0.462
E       2.527   1.594   2.557   0.937   2.821   1.374   1.398   3.132
A(I)    0.625   0.395   0.632   0.232   0.698   0.340   0.346   0.775

H SQ    0.391   0.156   0.400   0.054   0.487   0.115   0.120   0.600
E       2.611   1.639   2.640   0.953   2.936   1.402   1.435   3.271
A(I)    0.635   0.399   0.642   0.232   0.714   0.341   0.349   0.796

H SQ     .404   0.159   0.413   0.054   0.510   0.116   0.122   0.634
E       2.624   1.642   2.652   0.953   2.959   1.403   1.437   3.304
A(I)    0.637   0.398   0.644   0.231   0.718   0.341   0.349   0.802

H SQ    0.406   0.159   0.414   0.053   0.516   0.116   0.122   0.643
```

47

```
0807701    22  ENGINEERING MANAGEMENT    J. UTTERBACK          01100101   49
PREDICTED CORRELATION MATRIX

PRED R   SECTION  1

   VAR             ITEM01  ITEM02  ITEM03  ITEM04  ITEM05  ITEM06  ITEM07  ITEM08
 ITEM01
 ITEM02             0.000   0.254   0.410   0.147   0.458   0.217   0.222   0.511
 ITEM03             0.254   0.000   0.257   0.092   0.246   0.136   0.139   0.320
 ITEM04             0.410   0.257   0.000   0.149   0.462   0.219   0.225   0.514
 ITEM05             0.147   0.092   0.149   0.000   0.166   0.079   0.081   0.185
 ITEM06             0.458   0.246   0.462   0.166   0.000   0.245   0.251   0.576
 ITEM07             0.217   0.136   0.219   0.079   0.245   0.000   0.114   0.273
 ITEM08             0.222   0.139   0.225   0.081   0.251   0.114   0.000   0.280
                    0.511   0.320   0.514   0.185   0.576   0.273   0.280   0.000
RESIDUAL MATRIX

RESID    SECTION  1

   VAR             ITEM01  ITEM02  ITEM03  ITEM04  ITEM05  ITEM06  ITEM07  ITEM08
 ITEM01
 ITEM02             0.000  -0.080   0.064   0.050   0.069  -0.036  -0.106   0.037
 ITEM03            -0.080   0.000  -0.011  -0.003  -0.022   0.055   0.135  -0.074
 ITEM04             0.064  -0.011   0.000  -0.044   0.054   0.005  -0.163   0.097
 ITEM05             0.050  -0.003  -0.044   0.000  -0.146  -0.009   0.135   0.016
 ITEM06             0.069  -0.022   0.054  -0.146   0.000  -0.033   0.056   0.024
 ITEM07            -0.036   0.055   0.005  -0.009  -0.033   0.000   0.024  -0.008
 ITEM08            -0.106   0.135  -0.163   0.135   0.056   0.024   0.000  -0.085
                    0.037  -0.074   0.097   0.016   0.024  -0.008  -0.085   0.000




MEAN RESIDUAL EQUALS    0.0003SIGMA OF RESIDUALS IS    0.0737



NORTHWESTERN UNIVERSITY MAXIMUM LIKLIHOOD PROGRAM    TO EXTRACT ONE FACTOR
PRINCIPAL COMPONENTS COMMUNALITIES      0.5347     0.2117     0.5637     0.0761


                                        0.6272     0.1759     0.1488     0.6943
```

```
FINAL RESULTS          NUMBER OF ITERATIONS          RESIDUAL SUM OF SQUARES  [illegible]
COMMUNALITIES
     [illegible]     0.104983     0.514773     0.038107     0.573432     0.097232     0.069118     0.881983
SPECIFIC VARIANCES
     [illegible]     0.895017     0.485227     0.961893     0.426568     0.902768     0.930882     [illegible]
FACTOR LOADINGS
     0.677471     0.330125     0.717477     0.195019     0.757251     0.311820     0.262907     0.827020
        CORRELATION INPUT MATRIX AND RESIDUAL MATRIX
     1.000000     0.173544     0.674259     0.197008     0.526413     0.182995     0.118475     0.547861
    -0.050105     1.000000     0.245679     0.089189     0.244174     0.190447     0.274468     0.245146
    -0.011812     0.008842     1.000000     0.105181     0.515990     0.224017     0.061760     0.612928
     0.009274     0.026459    -0.031153     1.000000     0.020214     0.069472     0.210132     0.201634
     0.011396     0.014185    -0.027321    -0.123678     1.000000     0.211661     0.306830     0.603632
    -0.028254     0.087507     0.000243     0.010220    -0.024466     1.000000     0.143000     0.265117
    -0.059511     0.187677    -0.126867     0.166176     0.107746     0.061022     1.000000     0.194507
    -0.012421    -0.027874     0.019560     0.044485    -0.022631     0.007236    -0.022919     1.000000
CHI-SQUARE FOR LAWLEY TEST=  34.273454. DEG. FREEDOM IS       20


                                   WOLINS  INDEX OF SINGLE FACTOREDNESS
  SIGMA   =               0.3087   SIGMA OF RESIDUALS EQUALS    0.0494   EXPECTED  SIGMA              0.0214
   INDEX EQUALS     0.9021
```

On the Use of Both Pro and Con Items in Attitude Scales[1]

Donald T. Campbell

Northwestern University

Rorer's (1965) effective attack on "the great response style myth" may misleadingly lead future composers of attitude scales to revert to the less desirable practice of writing all items favoring one pole of the hypothesized attitude continuum, instead of having items stating both poles, balancing items so as to control for any direction-of-wording factor. What Rorer has validly done is to summarize the recurrent finding that this direction-of-wording component is not uniformly correlated across tests in such a way as to justify hypothesizing a general content-irrelevant response disposition. The term"acquiescence response set" has been misleading in this regard. He has also collated data showing that in not all scales is the direction-of-wording effect of large magnitude.

The fact remains, however, that when an attitude test constructor has hypothesized a bipolar attitude continuum, and has written items suitable for endorsement by persons at both poles (we can for convenience call these "pro" and "con" or "positive" and "negative") he will find that pro items correlate more highly among themselves, and that the con items correlate more highly among themselves, than the pro items correlate with the con ones. Similarly, in intercorrelations among a group of related attitudes where scales contain items worded in only one direction, it will be found that where the direction-of-wording is "the same" (e.g., where both sets of items

---

[1] The preparation of this paper was supported by Project C-998, Contract 3-20-001 with the Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

are worded in the "liberal" direction) the correlation will be higher than where the direction of wording differs (e.g., where one scale is worded in the "liberal" direction and one in the "conservative"). Most of the early work on response set in attitude scales was focussed on the methodological implications for interpreting correlation coefficients between scales where this direction-of-wording factor had not been held constant.

Lets take a concrete example. In *The Authoritarian Personality* study (Adorno, *et al.*, 1951), and particularly in the widely read preliminary statement (Frenkel-Brunswik, Sanford, & Levinson, 1948; 1965) it was argued that personality was more importantly related to prejudice than was political-economic conservativism, because the F scale (authoritarian personality trends) correlated with the E scale (ethnocentrism, anti-minority prejudice) .73, while the PEC scale (political-economic conservatism) correlated with the E scale but .50. But the fact that both the *E* and the *F* scales were worded solely in the conservative direction, while the *PEC* scale contained both liberal and conservative items, artifactually loaded the dice in favor of such a finding.[2]

For the effect of direction-of-wording upon the *F*-*E* correlation, we have some evidence. Table 1 presents the correlations among *F*, *E*, and *A* (The Taylor Manifest Anxiety Scale). (These data have been reassembled from Chapman & Campbell, 1959a, 1959b.) The major diagonal represents the reliability coefficients. Above the diagonal, the correlations are corrected for attenuation, below the diagonal are the uncorrected values. The data illustrate one of two more "successful" reversals of the F scale, since the correlation between the *F* pro and the *F* con is at least positive. However, it is low, and when corrected for attenuation reaches but .53. On the other hand,

Table 1

Direction-of-Wording in the Correlations Among the *F* Scale of Authoritarian Personality Trends, Ethnocentrism and Anxiety

(Values above diagonal corrected for attenuation)

	F_p	F_c	E_p	E_c	A_p	A_c
F_p	(.71)	.53	.76	.56	.41	.27
F_c	.29	(.42)	.36	.58	.12	.10
E_p	.57	.21	(.80)	.94	.13	.06
E_c	.39	.31	.70	(.69)	.05	.10
A_p	.32	.07	.11	.04	(.87)	.98
A_c	.21	.06	.05	.08	.84	(.85)

Ethnocentrism and Anxiety seem almost completely free of acquiescence bias, direction-of-wording making almost no difference, the correlations between positive and negative forms approaching unity (.94 and .98) when corrected for attenuation. But while the two forms of each seem equivalent, they are far from identical in the correlations they produce with F_p the original F scale. F_p correlates .57 with E_p , but only .39 with E_c the difference being significant at the .0002 level. The difference cannot be due solely to reliability differences, for it persists when corrected for attenuation (.76 vs. .56). Although Anxiety is still more immune to direction-of-wording effects than E, a parallel picture obtains here: F_p correlates .32 with A_p versus .21 with A_c (difference significant at $p < .01$), with the values corrected for attenuation becoming .41 and .27. Thus direction-of-wording is highly relevant to the magnitude of the correlation with the F scales even when only the "successful" reversals are used, the weak F_c being omitted from consideration.

Note that this effect is significantly present, even though no simple additive factorial component uniformly affects all the correlations in the matrix. Note that the correlation of E and A shifts from but .11 when both are positive to .04 or .05 when one or the other is negative. This is probably due to the fact that E and A have but small loadings on the direction-of-wording factor, if we take the magnitude of the positive-negative correlations, corrected for attenuation, as indicative. F, on the other hand, has a high loading by the same token.

While many attitude tests have been constructed in which balanced subsets of pro and con items have been used, it is rare to find them analyzed to demonstrate direction-of-wording effects in the intercorrelations between different attitudes. In an effort to provide another illustration, data reported earlier (Chapman & Campbell, 1957) have been reanalyzed for this purpose. (A few items stated in forced choice form have been omitted.) In the present analysis, all five levels of endorsement (Agree Strongly, Agree, Undecided, Disagree, Disagree Strongly) have been used in the scoring, and the Coefficient Alpha form of the Kuder-Richardson reliability (Cronbach, 1951) computed. The scales and numbers of items are as follows: F scale of authoritarian personality trends, F_p consisting of 15 of the original scale items (Adorno, *et al.*, 1950) and F_c consisting of 15 items worded in the non-authoritarian direction, revision following item analysis of a previous effort; SS scale of superior versus subordinate orientation, the items of which had gone through several prior utilizations with item analysis, SS_p and SS_c each containing 10 items; Al scale of alienation, developed from a number of scales used previously, 14 items each for Al_p and Al_c ; ID, identification with discipline, 7 items each for ID_p and ID_c ; NC, a scale reflecting non-cooperative or non social tendencies and preferences, used in this study for the first time, 13 items each for NC_p and NC_c , the direction of scoring and labeling for NC having been reversed from the earlier report to maximize positive correlations. The resulting matrix based upon data from 142 male undergraduates, enrolled in an introductory psychology course is reported in Table 2, with values above the diagonal corrected for attenuation by the reliabilities given in the diagonal.

Table 2 presents a striking contrast with Table 1, in that all scales show a strong direction-of-wording effect, none of the pro-con

Table 2

Intercorrelations Among Authoritarianism (F), Superior-Subordinate Orientation, Alienation, Identification with Discipline, and Non-cooperativeness as a Function of Pro or Con Wording of items. (Values above diagonal corrected for attenuation)

	Fp	Fc	SSp	SSc	Alp	Alc	IDp	IDc	NC$_p$	NC$_c$
Fp	(.62)	-.06	.54	-.28	.46	-.18	.46	-.16	.36	-.36
Fc	-.03	(.41)	-.47	.20	-.33	.17	.01	.25	-.09	.39
SSp	.29	-.21	(.47)	.53	.28	-.33	.75	-.10	-.29	-.54
SSc	-.14	.08	.23	(.42)	-.35	.46	.12	.40	-.22	.11
Alp	.28	-.20	.15	-.18	(.61)	.31	.21	-.51	-.24	-.57
Alc	-.10	.08	-.16	.21	.17	(.48)	-.57	.04	-.35	.23
IDp	.24	.00	.35	.05	.09	-.27	(.45)	.26	.15	-.15
IDc	-.09	.12	-.05	.19	-.27	.02	.13	(.54)	.22	.63
NC$_p$	.20	-.04	-.14	-.10	-.13	-.16	.07	.11	(.48)	.64
NC$_c$	-.22	.20	-.29	.05	-.35	.11	-.08	.36	.35	(.61)

correlations approaching unity when corrected by attenuation as did the E and A scales of Table 1. As in all studies, the F scale is reversed with poor success, direction-of-wording effects being in this case stronger than content. Again, looking only at the original F scale items, F_p , we find that direction-of-wording in the other scale produces correlation differences of .33 to .43 points, plus a change in the sign of the relationship. Perhaps because of a basic independence of content among the scales, or because of lack of conscientiousness on the part of the respondents (though they are from the same source as the respondents of Table 1), direction-of-wording produces not only highly significant differences in relationship, but also the most sizeable relationships.

Only one other study of attitudes both studying direction-of-wording and using a number of scales is known. Peabody (1961) reports a study in which all subjects were administered pro and con versions of the F scale, Dogmatism, Anti-Semitism, and Conservatism. While in the initial report, he did not present the full set of intercorrelations making possible analyses such as Tables 1 and 2 he has been kind enough to provide the present author with them and they are reported in Table 3. These data fully confirm direction-of-wording effects differentially important for different contents. The pro-con correlations corrected for attenuation are for his American sample .33, .37, .81, and 1.12 for F, Dogmatism, Anti-Semitism, and Conservatism, respectively, while for an English sample they run .60, .30, .96, and .94. These data show that F scale content (along with Dogmatism) is much more affected by direction-of-wording than is Anti-Semitism or Conservatism, in agreement with the parallel aspects of Table 1.

Table 3. Direction-of-Wording Effects in Peabody's Study of Authoritarianism (F) Dogmatism (D), Anti-Semitism (AS) and Conservatism (C). (Values above diagonal corrected for attenuation)

American Data (N = 88)

	F_p	F_c	D_p	D_c	AS_p	AS_c	C_p	C_c
F_p	(.69)	.33	.93	-.02	.59	.55	.02	-.28
F_c	.18	(.42)	-.20	1.01	.22	.52	-.19	.24
D_p	.61	-.10	(.62)	.37	.46	.18	.07	-.17
D_c	-.01	.47	.21	(.51)	.15	.49	-.33	.07
AS_p	.46	.13	.34	.10	(.87)	.81	.03	-.03
AS_c	.39	.29	.12	.30	.65	(.74)	-.08	-.05
C_p	.01	-.09	.04	-.17	.02	-.05	(.51)	1.12
C_c	-.18	.12	-.10	.04	-.02	-.03	.61	(.58)

British Data (N = 75)

	F_p	F_c	D_p	D_c	AS_p	AS_c	C_p	C_c
F_p	(.82)	.60	.84	.03	.68	.36	.09	-.05
F_c	.36	(.44)	.15	.92	.41	.60	-.08	.05
D_p	.63	.08	(.68)	.30	.60	.30	.16	-.14
D_c	.02	.49	.20	(.64)	.25	.51	-.11	-.01
AS_p	.54	.24	.44	.18	(.78)	.97	-.14	-.30
AS_c	.30	.37	.23	.38	.79	(.86)	-.16	-.10
C_p	.07	-.05	.12	-.08	-.11	-.13	(.79)	.94
C_c	-.04	.03	-.10	-.01	-.23	-.08	.74	(.78)

Again, if one looks only at the original F_p, its correlations are higher with "successfully" reversed scales such as AS and C as a function of the direction of wording of the latter. The pattern of F and AS for the British data is particularly striking. Dogmatism too shows a striking direction-of-wording effect in its relationship with Anti-Semitism, again even if only the original scale, D_p, is used.

Table 3

Direction-of-Wording Effects in Peabody's Study of Authoritarianism (F), Dogmatism (D), Anti-Semitism (AS) and Conservatism (C).

(Values above diagonal corrected for attenuation)

	Fp	**Fc**	**Dp**	**Dc**	**ASp**	**ASc**	**Cp**	**Cc**
Fp								
Fc								
Dp								
Dc								
ASp								
ASc								
Cp								
Cc								

Rorer himself has done an ambitious item-rewording study for the MMPI. (Rorer, 1965; Rorer & Goldberg, 1965a; 1965b). While he has not published an intercorrelational matrix examining the cross-trait effects of item direction, he has provided the present writer with some relevant computations made subsequent to the analyses which he has published. For certain scales of the MMPI which in original form had items keyed in both pro and con directions, he has provided 4 scores: original pro, original con, reversed wording pro, reversed wording con. While the resulting scales are short and hence produce lower and more inconsistent correlations than full length scales would, they are relevant for direction of differences even if not of absolute magnitude. Table 4 summarizes these results. In the full matrix there are 16 correlations between any two traits. These have been averaged pooling the correlations involving the same direction-of-wording, and all of those involving different directions, resulting in the two values shown for each interrelation shown in Table 4. While the pattern is somewhat uneven and the effect weak, shared direction-of-wording leads to higher correlations. If one takes as some minimum evidence that the same wording values average higher than the heterodirectional values then in 87 of the 90 cells a direction-of-wording effect is definitely present.

The direction-of-wording effects are small, as would be expected from the prior sampling of MMPI content (Campbell & Chapman, 1959b and Table 1 above). But they are definitely present. Since both directions of item wording work well, and since the magnitude of such effects cannot be assumed in advance of examination, these results are no reason to abandon the practice of having items worded in both directions, as indeed the original MMPI

Table 4

Direction-of-Wording Effects in Rorer's Study of MMPI Items

(S = same direction-of-wording, D = different. Males above diagonal, females below.)

	F		Hs		D		Hy		Pd		Pa		Pt		Sc		Ma		Dev	
	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D	S	D
F			.30	.30	.27	.25	.18	.13	.28	.24	.15	.11	.35	.33	.47	.41	.11	.08	.45	.37
Hs	.31	.29			.32	.30	.38	.33	.20	.19	.14	.12	.41	.39	.36	.35	.09	.08	.44	.41
D	.26	.24	.39	.35			.35	.22	.21	.12	.13	.07	.44	.40	.36	.31	-.05	-.13	.44	.36
Hy	.16	.11	.30	.26	.18	.11			.24	.12	.21	.10	.26	.21	.21	.17	.07	-.03	.29	.19
Pd	.25	.20	.19	.19	.23	.19	.20	.15			.15	.05	.22	.19	.32	.27	.23	.14	.33	.24
Pa	.03	.02	.03	.05	.06	.06	.09	.06	.09	.04			.10	.09	.15	.11	.04	-.04	.17	.09
Pt	.30	.30	.47	.44	.44	.40	.20	.14	.24	.20	.10	.08			.47	.45	.02	-.00	.55	.51
Sc	.40	.37	.44	.41	.35	.29	.22	.17	.29	.25	.10	.07	.50	.45			.13	.08	.54	.49
Ma	.06	.04	.05	.07	-.03	-.07	.09	.04	.14	.08	.06	.01	.02	-.01	.10	.06			.07	.00
Dev	.40	.35	.47	.43	.45	.38	.21	.14	.29	.22	.08	.05	.57	.51	.50	.44	.03	-.03		

scales did--although not in a systematically equated fashion.

Rorer's reports (1965; Rorer & Goldberg, 1965a, 1965b) are based upon intra-scale analyses, and they too show slight direction-of-wording effects. It should be noted, however, that two features in Rorer's setting lead to an exaggeration of the correlation across different-direction items within a scale and an underestimate of correlation between same-direction items. His four subscales of each dimension, Original pro, Original Con, Reversed-Pro, Reversed-Con, were administered in two testings, two weeks apart. *Op* and *Oc* were administered together, as were *Rp* and *Rc*. As is well known, measures given in the same administration correlate more highly than they do when separated by an interval. Thus $r_{O_pO_c}$ and $r_{R_pR_c}$ are higher than they would otherwise be, or than they should be relative to the other four correlations. Another irrelevant source of higher correlations comes from the sharing of specific illustrative content. Since *Rp* items are specific rewordings of *Oc* items, and since *Rc* items are specific rewordings of *Op* items, $r_{O_pR_c}$ and $r_{O_cR_p}$ are enhanced over what they would be had they been based as the other correlations are on separate samples of content. This would be true of any study, but is undoubtedly particularly true of Rorer's, where on the second testing the respondents must have felt a particular challenge to be consistent and to carefully decode the often awkward wording.

The awkwardness and the challenge come from Rorer's giving priority to exact logical contradictoriness over simplicity of wording or ease of comprehension. A few examples will indicate his vigilance in this regard:

Original	Reversed
I have a good appetite.	I do not have a good appetite (or it is about average).
My father was a good man.	My father was not a good man (or he was about average).
I seldom or never have dizzy spells.	It is not true that I seldom or never have dizzy spells.
I am not easily angered.	It is not true that I am not easily angered.
I daydream very little.	I daydream more than a very little.
It does not bother me particularly to see animals suffer.	It is not true that it does not bother me particularly to see animals suffer.

This awkwardness is so great that it would probably not have been feasible, in terms of student rapport, to have presented the reversed form as the first administration.

Thus for one reason the $r_{O_pO_c}$ and $r_{R_pR_c}$ values are too high, for another the $r_{O_pR_c}$ and $r_{O_cR_p}$ values. The values thus distorted are all values which the direction-of-wording factor reduces, and the estimates for that factor are thus made too small. Correspondingly, the trait factor is overestimated. Note that in most of the F scale rewording studies, and for the practical needs of attitude testers, the pro and con subsets share no specific item content, and if direct reversal has been employed, only one item of any given pro-con reversal pair is employed. To present a sample of what such a picture would look like for Rorer's data, Table 5 is presented, making use only of the pro and con subsets among the original items. Instability due to small numbers of items is a problem but the picture is still a clear one of response bias effects.

Table 5

Intercorrelations Among Pro and Con Subsets of Original MMPI Scales in Rorer's Study

(Males above diagonal, females below)

		F		Hs		D		Hy		Pd		Pa		Pt		Sc		Ma		Dev	
		P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C
F	P		.20	.56	.48	.53	-.13	.55	-.12	.72	-.18	.72	-.24	.63	.28	.72	.20	.52	-.15	.71	.16
	C	.21		.15	.21	.51	.23	.35	-.02	.34	.13	.21	-.09	.34	.37	.43	.62	.17	-.01	.31	.55
Hs	P	.65	.10		.61	.47	.12	.65	.12	.39	-.08	.62	-.17	.49	.36	.56	.12	.32	-.06	.51	.24
	C	.62	.21	.62		.57	.22	.64	.34	.43	.00	.51	-.09	.55	.56	.62	.30	.26	-.17	.50	.58
D	P	.50	.30	.51	.54		.19	.78	-.11	.62	-.09	.61	-.31	.82	.54	.81	.53	.39	-.32	.74	.56
	C	.19	.29	.32	.39	.19		.13	.50	-.11	.29	-.01	.29	-.03	.39	.01	.16	-.40	-.08	-.18	.51
Hy	P	.54	.20	.67	.60	.55	.19		.01	.55	-.10	.60	-.23	.73	.50	.79	.32	.40	-.20	.65	.46
	C	.04	.06	.06	.23	-.23	.32	-.01		-.17	.48	-.13	.64	-.28	.26	-.14	-.02	-.34	.36	-.41	.39
Pd	P	.70	.25	.49	.60	.65	.16	.63	-.12		-.06	.64	-.30	.74	.43	.80	.46	.64	-.16	.80	.33
	C	.02	.10	-.08	-.02	-.09	.35	-.07	.47	-.02		-.17	.36	-.22	.05	-.16	.14	-.20	.28	-.23	.31
Pa	P	.59	.20	.47	.54	.68	.17	.51	-.18	.72	-.13		-.37	.75	.38	.78	.23	.52	-.22	.77	.26
	C	-.40	-.13	-.25	-.39	-.45	-.08	-.29	.40	-.47	.12	-.37		-.43	-.05	.33	-.12	-.44	.17	-.55	.24
Pt	P	.56	.21	.52	.58	.81	.10	.68	-.33	.77	-.20	.78	-.51		.52	.88	.43	.60	-.39	.89	.44
	C	.43	.36	.40	.65	.60	.42	.53	.05	.54	.03	.51	-.31	.61		.52	.41	.18	-.24	.44	.62
Sc	P	.65	.24	.49	.59	.72	.08	.68	-.22	.83	-.06	.76	-.51	.86	.55		.44	.57	-.32	.86	.48
	C	.49	.53	.41	.49	.35	.30	.33	.11	.40	.05	.31	-.09	.34	.38	.36		.22	-.02	.40	.58
Ma	P	.52	.04	.34	.43	.46	-.14	.46	-.22	.68	-.14	.61	-.52	.67	.34	.77	.17		-.03	.73	-.05
	C	-.18	-.37	-.19	-.31	-.49	-.13	-.24	.42	-.38	.34	-.40	.39	-.52	-.40	-.39	-.23	-.18		-.36	-.16
Dev	P	.68	.19	.56	.62	.77	.04	.61	-.37	.80	-.23	.79	-.60	.91	.53	.88	.33	.78	-.50		.27
	C	.30	.57	.29	.56	.40	.59	.39	.26	.34	.26	.28	-.14	.35	.60	.32	.46	.01	-.29	.24	

Some of these are quite marked. *Pa* seems particularly susceptible, for note the negative correlation of -.37 between the Pro and Con subsets of *Pa* items, replicated in both male and female samples. *Pd* seems another weak scale, with pro-con subset correlations of -.06 and -.02; *Hy* has .01 and -.01; *Ma* has -.03 and -.18. If one looks then at the intercorrelations among these direction-of-wording prone scales, the pattern of wording-effects on inter-trait correlations is as dramatic as that provided in the *F* scale literature, providing correlation differences of over 1.10 correlation points (as from .72 to -.47). Table 6 pulls out of Table 5 the data for four highly susceptible scales.

In many of the instances of Table 6, as in the *.60* -.23/-.13 .64 pattern of males Pa · Hy, the wording effect is clearly symmetrical, the pro items and the con items both hanging together across traits to enhance correlations. More characteristic of the matrix, however, is the finding that the pro items (items scored true for diagnostic category) carry much more covariance with them than do items scored con, the con-con cell often being lower than one of the pro-con cells. This fact probably shows some sort of confound between direction of item wording and type of symptom content used in the item, and this may too, be involved in part in the dramatic effects of Table 6. Since in the MMPI, no planning on direction of item wording was done, it seems likely that some systematic factors determined which items were worded in what way. But as Table 4 shows, even where the content sampling problem is clearly removed, in that the reversed item data are employed, a persistent direction-of-wording effect is present, even though not strong. Certainly the data do not justify neglect of the problem.

Table 6

Intercorrelations as a Function of Direction-of-Wording for

Four Susceptible MMPI Scales

		Hy		Pd		Pa		Ma	
		P	C	P	C	P	C	P	C
Hy	P		-.01	.55	-.10	.60	-.23	.40	-.20
	C	-.01		-.17	.48	-.13	.64	-.34	.36
Pd	P	.63	-.12		-.06	.64	-.30	.64	-.16
	C	-.07	.47	-.02		-.17	.36	-.20	.28
Pa	P	.51	-.18	.72	-.13		-.37	.52	-.22
	C	-.29	.40	-.47	.12	-.37		-.44	.17
Ma	P	.46	-.22	.68	-.14	.61	-.52		-.03
	C	-.24	.42	-.38	.34	-.40	.39	-.18	

The present writer knows of no study of attitude scales which has looked for direction-of-wording effects in the relationship among scales or subscales and failed to find them. Such matrices when examined in unpublished dissertation research at Northwestern regularly confirm them. Rorer reports no exceptions, the stated conclusions of his review notwithstanding. Certainly these direction-of-wording effects are less for some contents than others, with MMPI content (as in the Taylor Manifest Anxiety Scale, A in Table 1, above) and specific ethnic group content (as the _E_ scale of Table 1 and Anti-Semitism of Table 3) are relatively immune, but still significantly affected when the correlation with other variables is examined. For the average attitude scale direction-of-wording effects are almost certainly present to a significant degree. Certainly they should be controlled for rather than assumed absent.

This conclusion differs so greatly in tone from that of Rorer (1965) that the focus of disagreement had best be clarified. One of his conclusions we can agree with unambiguously: there seems to be no single strong direction-of-wording effect operating as a consistent personality factor across all tests and contents. We can agree in part even with his exoneration of the MMPI scales of large direction-of-wording components. We strongly disagree, however, with any conclusion that direction-of-wording could have been neglected without jeopardy.

Rorer's most untenable conclusion is as to the absence of evidence of response-set in the _F_ scale, which he gives as clean a bill of health as he does the MMPI. The research data are quite in contrast. Every study made of the _F_ scale has found impressive evidence of direction-of-wording effects.

Even the paper of Christie, et al., purportedly a successful reversal, finds a correlation of but .05 between the positive and negative forms for college students, the population most comparable to those of other studies. Rorer exonerates the MMPI studies because most reversals have been successful. The F scale he exonerates because none have been successful. Rorer is able to reject these studies by requiring that each original item be paired with a reversal which is its logical contrary--a requirement which is essentially irrelevant. A review of the history of attitude test construction on this point may help clarify the essential goals that those who wrote F con items had, or should have had in mind.

The attitude test constructor came naturally to bipolar sets of items within an attitude scale. In the older tradition (Allport & Hartman, 1925, Thurstone & Chave, 1929) there was an effort to provide items stating each position along the entire length of an attitude dimension. When this tradition was modified for Likert-type scoring or a Guttman scale analysis, non-monotonic items, particularly frequent in the neutral zone of the older scales, were eliminated, and there remained stronger and weaker statements of the pro end of the scale plus strong and weak statements of the con position. The major methodological touchstone for the authors of the authoritarian personality study was Likert, and on this point he says:

> To avoid any space error or any tendency to a stereotyped response, it seems desirable to have the different statements so worded that about one half of them have one end of the attitude continuum corresponding to the left or upper part of the reaction alternatives and the other half have the same end of the attitude continuum corresponding to the right or lower part of the reaction alternatives. For example, about one-half of the statements in the Internationalism scale have the international extreme corresponding with 'Strongly Approve' while the other half has it corresponding with 'Strongly Disapprove.' These two kinds of statements

> ought to be distributed throughout the attitude test in a chance or haphazard manner. (Likert, 1932, p. 46)

The wisdom of this dominant tradition now seems clear. The decision of Adorno, et al. (1950) to deviate from it in their Anti-Semitism and Ethnocentrism scales is the historical anomaly, which through internal and cross-scale item selection eliminated the few democratic items from the F scale. When researchers sought to correct the F scale in this regard, their goal was to write items stating the position of persons at the nonauthoritarian end of the scale. They did this, for reasons of convenience, by "reversing" half of the items in the original F. The goal of such reversal was not at all to get an item with the same frequency of authoritarian answers as the original, rather, it was simply one of getting an item for which agreement would symptomize on the average low F tendencies. In this, there was no requirement to achieve a logical contrary, particularly at the expense of the awkward phraseology illustrated in the sample of Rorer's MMPI reversals above. To reject such reversals merely because one can in consistency reject both (or, more rarely, accept both) is as unreasonable as to require logical and response identity between two items worded in the same direction. Item reversal began as but a technique for getting some con items. Rorer has elevated it into the sole criterion of direction-of-wording effects, rejecting prior studies on this ground.

That the original authors of the F scale found it conceivable that there were such items is shown by their inclusion of four out of thirty-eight low F items in the original item pool--items that dropped out in the item analysis. That some F scale items, original and reversed, could have been selected through item analysis with a balanced *E* scale, has also been demonstrated (Chapman & Campbell, 1959a). Many *F* scale items do not survive this criterion, however.

To exonerate the *F* scale of direction-of-wording effects it would seem necessary to put it to the test and show the absence of such effects. Rorer cites no study doing this for the *F* scale.

In the tradition here advocated, the less the success of reversals, the greater the direction-of-wording effects--these to be manifest not only in interval analyses of reversal, but also, as in Tables 1, 2, and 3, through the relationship of either form to other variables as a function of the direction of the wording of these other variables. Ambiguity of item meaning would be expected to show up not only in the difficulty to writing item reversals, but also in the difficulty of content--relevant responding, leaving the covariance of responses dominated by other than content considerations. To reiterate again, note the contrast in type of evidence between Rorer's exoneration of the MMPI and *F* scale. The MMPI is exonerated because all item reversal studies have been successful, have found direction-of-wording effects. The *F* scale has been exonerated because all efforts to reverse it have failed. It is here contended that until successful reversal studies are reported, these should be interpreted as originally offered, as showing preponderant direction-of-wording effects. But even disregarding the reversal data per se, evidence, as in Tables 1, 2, and 3, shows strong direction-of-wording effects in the relationship between the original *F* scale and other measures presumably successfully reversed.

Summary

Direction-of-wording effects are demonstrated in the correlations between attitude scales. Thus although Ethnocentrism and Anxiety scales are successfully "reversed" and thus appear free of direction-bias, their correlations with the original F scale are significantly affected by direction-of-wording. All available studies are confirmatory. Rorer's MMPI item reversal study shows some direction-of-wording effects on the interrelations among all variables, particularly strongly for scales such as Hy, Pd, Pa, and Ma.

In criticism of Rorer's conclusions, it is noted that he exonerates the MMPI because, when studied, direction-of-wording effects are slight, while he exonerates the F scale on the grounds of inadequacy of the studies, although all have shown strong effects. His requirement of logical contradictions in item reversal studies is discussed as irrelevant to the use of reversal as an optional means to achieving a pool of items worded in the con direction. The tradition of including items worded in advocacy of both poles of an attitude continuum is endorsed.

References

Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J., & Sanford, R. N. The authoritarian personality. New York: Harper, 1950.

Allport, F. H., & Hartman, D. A. The measurement and motivation of an atypical opinion in a certain group. American Political Science Review, 1925, 19, 735-760.

Chapman, L. J., & Bock, R. D. Components of variance due to acquiescence and content in the F scale measure of authoritarianism. Psychological Bulletin, 1958, 55, 328-333.

Chapman, L. J., & Campbell, D. T. Response set in the F scale. Journal of Abnormal and Social Psychology, 1957, 54, 129-132.

Chapman, L. J., & Campbell, D. T. An attempt to predict the performance of three-man teams from attitude measures. Journal of Social Psychology, 1957, 46, 277-286.

Chapman, L. J., & Campbell, D. T. The effect of acquiescence response-set upon relationships among F scale, ethnocentrism, and intelligence. Sociometry, 1959, 22, 153-161. (a)

Chapman, L. J., & Campbell, D. T. Absence of acquiescence response set in the Taylor Manifest Anxiety scale. Journal of Consulting Psychology, 1959, 23, 465-466.

Christie, R., Havel, Joan, & Seidenberg, B. Is the F scale irreversible? Journal of Abnormal and Social Psychology, 1958, 56, 143-159.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. The antidemocratic personality. In T. M. Newcomb & E. L. Hartley, (eds.), Readings in Social Psychology. New York: Holt, 1947, 531-541; Revised Edition, G. E. Swanson, T. M. Newcomb, & E. L. Hartley, 1952, pp. 612-622; Third Edition, E. E. Maccoby, T. M. Newcomb, & E. L. Hartley, 1958, pp. 636-646; retitled, The authoritarian personality, H. Proshansky, & B. Seidenberg, (eds.), Basic studies in social psychology, New York: Holt, 1965, pp. 670-679.

Likert, R. A technique for the measurement of attitudes. Archives of Psychology, 1932, 22, No. 140, 1-55.

Peabody, D. Attitude content and agreement set in scales of authoritarianism, dogmatism, antisemitism, and economic conservatism. Journal of Abnormal and Social Psychology, 1961, 63, 1-11.

Rorer, L. G. The great response-style myth. Psychological Bulletin, 1965, 63, 129-156.

Rorer, L. G., & Goldberg, L. R. Acquiescence in the MMPI? Educational and Psychological Measurement, 1965, 25, in press (a)

Rorer, L. G., & Goldberg, L. R. Acquiescence and the vanishing variance component. Journal of Applied Psychology, 1965, 49, in press. (b)

Thurstone, L. L., & Chave, E. J. The measurement of attitude. Univ. of Chicago Press, 1929.

Footnotes

[1] The preparation of this paper was supported by Project C 998, Contract 3-20-001 with the Office of Education, U.S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

[2] For this particular comparison of coefficients three other irrelevant sources of correlation were also involved. The *F* scale had 45 items, while the *PEC* scale had but 10. The *F* scale had been built up on the basis of several rounds of item analyses in which correlation with the *E* scale was the main selection criterion, while the *PEC* scale had no such advantage. In these ways the dice were loaded in favor of a Freudian and against a Marxist interpretation of prejudice. However, some of the *PEC* items had a personality flavor irrelevant to political-economic conservativism, and this factor may have been dominant, for Peabody (1961) finds almost no correlation between *PEC* and either *F* or *E* when more clearly relevant items are used.

Some Simplex Matrices Based on High School Grades and Absences[1]

Richard M. Rozelle and Donald T. Campbell

---

[1] This study was supported by Project C998, Contract 3-20-001, with the Office of Education, U. S. Department of Health, Education, and Welfare, under provisions of Title VII of the National Defense Education Act.

In the course of another study[2] we have generated a number of correlational matrices which conform to the general simplex pattern. Since the correlations are based upon very large numbers of cases, they have a stability which might make them useful in trying out alternative modes of factoring simplexes. They are being distributed for their possible value to research workers in this area.

The data consisted of grade and attendance records of students who attended Evanston Township High School, Evanston, Illinois. The records were from students who had attended the school from 1959 through 1962. There were six working periods with the first three marking periods designated as the first semester and the second three marking periods designated as the second semester. Simultaneously with the third-period grade a first semester grade was provided. At the time of the sixth marking period, a second semester and final grade were reported. During the fifth marking period a special "Spring Examination" was given, which was recorded separately and which also contributed to the grade for the fifth marking period and the second semester grade. These later four variables are also included, and might be useful in explaining systematic departures of the matrix of six marking periods from the pattern of a pure simplex.

---

[2] Richard M. Rozelle. An exploration of two quasi-experimental designs: the cross-lagged panel correlation and the multiple time series. M. S. Thesis, Northwestern University, 1965, Multilithed copies available.

The school records contained one computer card for each student for each course, on which the grade and attendance records were recorded. It is important to emphasize that the unit of analysis was the information given on each computer card and not the individual student. Thus, the number of computer cards was the N used, not the number of students. Because a number of different courses were used, most of the students were represented in more than one analysis, and represented more than once in the pooled analysis.

A total of five different course combinations were analyzed. These were:

1. First Year Foreign Languages
2. Second and Third Year Foreign Languages
3. Combined Studies, English, and Social Studies
4. Science Courses
5. Mathematics

A sixth matrix based on the pooling of all the data from the five course divisions was also computed. In all analyses, anomalous grades, "opportunity" courses, and ungraded students (5) have been eliminated.

The correlation matrices are organized so that examination of the first six variables (i.e., the individual marking period grades) should reveal a simplex pattern. The correlations of grades for first year foreign language courses form a perfect simplex pattern, at least as far as uniform decrease of correlation magnitude is concerned. Each of the remaining five grade-correlation matrices approach the form of a pure simplex but deviations from this pattern occur in the sixth grading period.

This higher correlation of the sixth marking period grade with grades of earlier marking periods is probably due to a tendency to assimilate this grade to the final grade, assigned at the same time, i. e., to make it more than a report on the work of that marking period. The tendency for the correlation between adjacent marking periods to steadily increase probably reflects a similar carry-over of impressions from previous periods in evaluating any given marking period.

The matrices for absences are also reported with the first six variables representing absences for each of the six individual marking periods. The first semester absence records consisted of the sum of the absences of the first three marking periods and the second semester absence records consisted of the sum of the absences of the last three marking periods. Due to skewness of distributions, pile-ups of cases at zero absences, the appropriateness of these data for factoring is much less than is the case for grades. There are a few reversals of simplex rank in each matrix, but these are less systematic than in the grade matrices.

Table 1. Grades in First Year Foreign Languages (N=1087)

	G_1	G_2	G_3	G_4	G_5	G_6	1st Sem. G	2nd Sem. G	Final G	Spring Exam
G_1	1.000									
G_2	.780	1.000								
G_3	.678	.782	1.000							
G_4	.652	.724	.790	1.000						
G_5	.594	.687	.733	.789	1.000					
G_6	.545	.646	.701	.766	.797	1.000				
1st Sem.	.873	.915	.870	.761	.717	.667	1.000			
2nd Sem.	.627	.729	.782	.866	.889	.914	.752	1.000		
Final	.683	.777	.815	.865	.875	.892	.810	.977	1.000	
Spring Exam	.601	.690	.708	.762	.760	.737	.712	.849	.845	1.000

Table 2. Grades in Second and Third Year Foreign Languages (N=911)

	G_1	G_2	G_3	G_4	G_5	G_6	1st Sem. G	2nd Sem. G	Final G	Spring Exam
G_1	1.000									
G_2	.737	1,000								
G_3	.715	.758	1.000							
G_4	.693	.744	.789	1.000						
G_5	.669	.715	.732	.774	1.000					
G_6	.661	.719	.731	.760	.781	1.000				
1st Sem. G	.866	.889	.881	790	.757	.749	1.000			
2nd Sem. G	.723	.772	.796	.853	.870	.924	.816	1.000		
Final G	.739	.788	.812	.854	.869	.917	.837	.993	1.000	
Spring Exam	.692	.700	.717	.738	.725	.737	.753	.839	.835	1,000

Table 3. Grades in Combined Studies, English, and Social Studies (N=5972)

	G_1	G_2	G_3	G_4	G_5	G_6	1st Sem. G	2nd Sem. G	Final G	Spring Exam
G_1	1.000									
G_2	.766	1.000								
G_3	.734	.775	1.000							
G_4	.722	.744	.775	1.000						
G_5	.700	.738	.754	.778	1.000					
G_6	.713	.744	.760	.764	.783	1.000				
1st Sem. G	.872	.903	.885	.792	.778	.785	1.000			
2nd Sem. G	.763	.790	.810	.857	.870	.919	.837	1.000		
Final G	.777	.806	.823	.856	.870	.915	.855	.992	1.000	
Spring Exam	.677	.682	.698	.685	.706	.709	.724	.813	.810	1.000

Table 4. Grades in Science Courses (N=1612)

	G_1	G_2	G_3	G_4	G_5	G_6	1st Sem. G	2nd Sem. G	Final G	Spring Exam.
G_1	1.000									
G_2	.769	1.000								
G_3	.769	.796	1.000							
G_4	.702	.757	.789	1.000						
G_5	.702	.752	.767	.796	1.000					
G_6	.698	.762	.776	.785	.806	1,000				
1st Sem. G	.883	.905	.910	.796	.786	.791	1.000			
2nd Sem. G	.752	.813	.829	.874	.888	.926	.847	1.000		
Final G	.775	.831	.848	.874	.882	.921	.874	.989	1,000	
Spring Exam	.687	.696	.699	.687	.719	.692	.730	.807	.800	1,000

Table 5. Grades in Mathematics (N=2748

	G_1	G_2	G_3	G_4	G_5	G_6	1st Sem. G	2nd Sem. G	Final G	Spring Exam
G_1	1.000									
G_2	.718	1.000								
G_3	.706	.737	1.000							
G_4	.655	.707	.750	1.000						
G_5	.628	.676	.708	.744	1.000					
G_6	.634	.679	.726	.746	.758	1.000				
1st Sem. G	.866	.882	.880	.759	.721	.732	1.000			
2nd Sem. G	.683	.734	.780	.841	.858	.919	.788	1.000		
Final G	.713	.763	.804	.841	.852	.908	.822	.988	1.000	
Spring Exam.	.598	.635	.684	.688	.719	.703	.685	.825	.818	1.000

Table 6. Grades in All Courses Combined (N=12330)

	G_1	G_2	G_3	G_4	G_5	G_6	1st Sem. G	2nd Sem. G	Final G	Spring Exam
G_1	1.000									
G_2	.760	1.000								
G_3	.729	.772	1.000							
G_4	.700	.740	.776	1.000						
G_5	.677	.725	.745	.777	1.000					
G_6	.676	.724	.751	.765	.784	1.000				
1st Sem. G	.875	.900	.887	.786	.764	.763	1.000			
2nd Sem. G	.731	.777	.805	.858	.873	.921	.821	1.000		
Final G	.754	.799	.823	.858	.870	.913	.848	.990	1.000	
Spring Exam	.660	.680	.701	.700	.720	.712	.721	.822	.817	1.000

Table 7. Absences in First Year Foreign Languages (N=1087)

	A_1	A_2	A_3	A_4	A_5	A_6	1st Sem. A	2nd Sem. A
A_1	1.000							
A_2	.389	1.000						
A_3	.391	.535	1.000					
A_4	.257	.396	.402	1.000				
A_5	.278	.382	.395	.514	1.000			
A_6	.244	.323	.360	.374	.469	1.000		
1st Sem. A	.769	.778	.822	.441	.442	.389	1.000	
2nd Sem. A	..326	.461	.454	.799	.814	.777	.532	1.000

Table 8. Absences in Second and Third Year Foreign Languages (N=911)

	A_1	A_2	A_3	A_4	A_5	A_6	1st Sem. A	2nd Sem. A
A_1	1.000							
A_2	.283	1.000						
A_3	.268	.247	1.000					
A_4	.215	.242	.259	1.000				
A_5	.251	.243	.304	.309	1.000			
A_6	.175	.151	.239	.265	.306	1.000		
1st Sem. A	.690	.685	..762	.335	.376	.269	1.000	
2nd Sem. A	.286	.283	.360	.732	.666	.771	.438	1.000

Table 9. Absences in Combined Studies, English, and Social Studies (N=5972)

	A_1	A_2	A_3	A_4	A_5	A_6	1st Sem. A	2nd Sem. A
A_1	1.000							
A_2	.385	1.000						
A_3	.305	.382	1.000					
A_4	.238	.306	.377	1.000				
A_5	.310	.328	.344	.391	1.000			
A_6	.283	.325	.334	.348	.428	1.000		
1st Sem. A	.725	.741	.794	.413	.434	.416	1.000	
2nd Sem. A	.357	.414	.458	.777	.759	.771	.546	1.000

Table 10. Absences in Science Courses (N=1612)

	A_1	A_2	A_3	A_4	A_5	A_6	1st Sem. A	2nd Sem. A
A_1	1.000							
A_2	.258	1.000						
A_3	.210	.228	1.000					
A_4	.212	.321	.251	1.000				
A_5	.231	.283	.276	.366	1.000			
A_6	.235	.288	.227	.267	.343	1.000		
1st Sem. A	.679	.635	.765	.367	.375	.351	1.000	
2nd Sem. A	.305	.402	.338	.754	.731	.737	.490	1.000

Table 11. Absences in Mathematics (N=2748)

	A_1	A_2	A_3	A_4	A_5	A_6	1st Sem. A	2nd Sem. A
A_1	1.000							
A_2	.412	1.000						
A_3	.362	.376	1.000					
A_4	.264	.343	.314	1.000				
A_5	.279	.296	.359	.332	1.000			
A_6	.252	.294	.340	.323	.333	1.000		
1st Sem. A	.763	.736	.797	.398	.409	.388	1.000	
2nd Sem. A	.355	.420	.452	.767	.716	.745	.535	1.000

Table 12. Absences in All Courses Combined (N=12330)

	A_1	A_2	A_3	A_4	A_5	A_6	1st Sem. A	2nd Sem. A
A_1	1.000							
A_2	.369	1.000						
A_3	.312	.366	1.000					
A_4	.241	.321	.340	1.000				
A_5	.286	.316	.341	.383	1.000			
A_6	.257	.299	.315	.328	.391	1.000		
1st Sem. A	.732	.728	.791	.401	.419	.387	1.000	
2nd Sem. A	.342	.410	.436	.771	.746	.761	.528	1.000

To appear, abbreviated, in the Journal of Personality and Social Psychology, 1965

Zero-Point Scaling as Affected by Social Object, Scaling Method, and Context

Jack M. Hicks
Wake Forest College
and Donald T. Campbell
Northwestern University

The desire to achieve a meaningful zero point for psychological measures is a recurrent one. Recently Thurstone and Jones (1959) have provided a promising method within the tradition of the law of comparative judgment. This paper extends the exploration of their method by using a wider range of social objects scaled, by using several different scaling methods, and most importantly, by examining the extent to which the zero-based scale values for a given object remain the same when the scaling is done in the context of sets of objects varying in scale value.

Fundamental to the technique of Thustone and Jones is the demonstration that the scale value of two stimuli when judged as a combination be consistently greater than the scale values of either stimulus when judged alone. This established, it is possible to apply the mathematical concept of additivity and infer a rational zero point. By assuming that the scale value of a combination of two stimuli, AB, is equivalent to the sum of the scale values of the individual stimuli, A and B, they take the combined distances of A and B from AB as the distance of AB to the

---

[1] Supported in part by U.S. Office of Education Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

zero point. That is, if AB is two units above A and three units above B, the rational zero point may be deduced to be five units below AB. With N basic stimuli, N(N-1)/2 separate zero-point estimations are possible. The crucial test of additivity concerns the variability among these zero-point determinations. The less the variability, the more convincing is the mean of the zero-point estimations as representative of a rational origin. Thurstone and Jones (1959) and a follow-up by Shuford, Jones and Bock (1960) have demonstrated the applicability of this requirement using the method of paired comparisons.

With regard to the choice of social objects scaled, Thurstone and Jones (1959) provide a limitedand atypical sampling. They asked for paired comparisons of five well-known consumer articles in terms of preferences as "birthday gifts which may not be sold." To achieve subjectivity of judgment, the respondents were instructed to ignore the objective judgmental criterion of resale value, and they did indeed obtain a negligible relationship between subjective scale values and monetary values of the items. While it would seem, therefore, that the objective criterion of monetary value was essentially ruled out, it could be argued that the mere availability of well-known objective standards disqualifies consumer articles as adequate tests of "subjective additivity." Therefore, in the present study two stimulus dimensions with lesser degrees of association with objective "values" have been chosen. These are the judged seriousness of traffic violations and the judged bizarreness of behavior symptoms. Birthday gifts were also included for replicative purposes.

It was hypothesized that traffic violations should be subjectively additive. Due to the commonplace occurrence of multiple violations committed by the same person, it becomes necessary for one to "add up" the component violations in order to estimate the total seriousness of the offenses. In this respect, traffic violations are seen as similar to birthday gifts, which also occur in multiples for a single individual. Violations differ from gifts, however, in the absence of a single, relatively unambiguous objective criterion, in that the public is less generally exposed to consistent objective outcomes associated with traffic violations than with birthday gifts. Practically all consumer articles are associated with a purchase according to a relatively stable retail price, whereas few traffic violations result in fines or even apprehension. Even when fines do occur, they may vary radically according to conditions, such as degree of traffic hazard involved. Thus the "objective" criteria of amount of fine or traffic hazard might be less readily brought to bear upon the judgment of the seriousness of a traffic violation than the criterion of resale price in the case of gifts. In the absence of a clear-cut objective criterion, past social experience is expected to provide a basis for judgments of a predominantly subjective nature. Assuming stimuli with objective judgmental criteria to be more readily added than subjective stimuli, somewhat less evidence of additivity is expected for traffic violations than for gifts.

The stimulus dimension of "bizarreness of behavior symptoms" was not expected to provide subjective additivity. This dimension appeared to have neither socially anchored subjective additive properties nor clear-cut objective criteria. In lieu of additivity, an "averaging" effect was predicted for the behavioral bizarreness dimension. Pairs of behavior statements might be judged as small samples of a total universe of potential information about an individual, rather than as exhaustive. Thus an "extremely bizarre" statement combined with a "mildly bizarre" statement might be judged as representative of a "moderately bizarre" personality. The phenomenon of averaging of stimulus combinations has also been suggested by Osgood and Sebeok (1954) with regard to adjective-noun combinations in their discussion of "laws of word mixture."

A second major objective of this study concerned the effects of stimulus context upon zero-point scaling. This was essentially an attempt to determine the relative positions of zero points independently derived from "high" and "low" contexts of a common dimension. Ideally, such zero points would be the same, and the resultant absolute scale values for a given stimulus would be independent of the context in which it was judged. In general, however, the experimental literature suggests a context effect. The dependence of specific scale values upon stimulus context has been a universal finding in psychophysics as well as in the areas of affective, aesthetic and social judgments (e.g., Helson, 1947; 1948; 1959). A "tracer" stimulus common to two contexts will have a lower scale value when judged among stimuli higher on the dimension than when judged among lower stimuli (e.g., Campbell, Lewis, & Hunt, 1958; Krantz & Campbell, 1961). Therefore, in terms of the present investigation, a given tracer stimulus should decrease in scale value (i.e., scale distance from zero) as the context increases in mean magnitude.

In anticipation, it may be noted that we do find such context effects, and in very strong degree. On the one hand, these effects may reflect genuine changes in subjective valuation as a function of context. On the other hand, they may be but methodological artifacts due to the untenability of a basic assumption of all applicable versions of Thurstone's law of comparative judgment. In moving from Thurstone's case 1 which holds for multiple judgments of a single person, to cases 2, 3, 4, and 5 which apply to pooled individuals, one presumably assumes that these individuals are homogeneous in their discriminatory performance. For those studies in which relative metric stimulus values is all that is needed, or all that can be checked, violations of the assumption introduce very minor problems. In zero-point scaling they have differential effects in differing contexts, producing a spurious context effect. In pursuance of this problem, a synthetic illustration of random error artifacts is provided, as well as a subsidiary analysis dividing judges into high and low discriminatory efficiency.

A third objective of this investigation was to determine the feasibility of categorical judgments methods for the Thurstone-Jones zero-point model. Although Rimoldi (1956) has used combined stimuli in the method of successive intervals, his study was not designed specifically to solve for a rational origin. In view of the greatly reduced labor involved in categorical assignments as opposed to paired comparisons, and the promising results reported by Jones (1957) in his reanalysis of the Rimoldi data for a rational origin, this additional exploration was undertaken. Moving to still less laborious procedures, approaching Thurstone's method of equal appearing intervals, an absolute judgments analysis or method of mean ratings was investigated.

Method

Subjects. Four hundred eighty-two freshman and sophomore college students of both sexes participated as judges. Three hundred seventeen of these judges were students at Wright Junior College, Chicago, and 165 Ss were enrolled at Northwestern University.

Materials. Stimulus materials were verbal statements of three types of objects: (1) consumer articles; (2) traffic violations; and (3) behavior symptoms. One item represented each of nine scale values of each stimulus group, and is identified in analysis by capital letters A through I. The letter A represents the "high" end of each dimension, and I the low end.

Consumer articles ranged along a dimension of "desirability as a birthday gift," as follows: (A) TV Set; (B) Record Player; (C) Camera; (D) Typewriter; (E) Transistor Radio; (F) Leather Binder; (G) Desk Lamp; (H) Pen and Pencil Set; (I) Dictionary. Items E,G, H,and I were borrowed from the Thurstone-Jones investigation and were ranked in accordance with those findings. The remaining articles were introduced to extend the range of stimuli, and are ranked according to relative retail price range.

Traffic violations ranged along a dimension of "seriousness of traffic offense," as follows: (A) Driving Under the Influence of Alcohol; (B) Leaving the Scene of an Accident; (C) Driving after License has been Revoked; (D) Driving 50 mph. in a 25 Mile Zone; (E) Running a Red Light; (F) Running a Stop Sign; (G) Going the Wrong Way on a One-Way Street; (H) Following a Car too Closely; and (I) Failing to Signal Before Turning. These items were borrowed from an unpublished study by Dr. Lawrence Ross. The behavior symptoms were taken from unpublished work by Dr. Jack Gilchrest, who had scaled them as to degree of "bizarreness and eccentricity." The following statements were used: (A) He tried to squeeze his head between the bars of his bed; (B) He kept hiding in dark corners and crying; (C) He said the green walls of his room were really purple; (D) He went on a number of hunger strikes; (E) He often said, "I wish I were dead"; (F) One day he brushed his teeth six times; (G) He seldom smiled; (H) He was found walking in his sleep; (I) He often complained of a backache.

Design. There were three basic experimental conditions: high context, low context, and neutral context. Items A through E were classified as high context values, and E through I as low context. The neutral condition consisted of alternate stimuli A, C, E, G, and I. Stimulus E was used as a tracer, appearing in all three contexts. Each of the three stimulus dimensions were divided in this manner into high, neutral, and low contexts, resulting in nine item sets. These nine sets were presented by both categorical assignment and paired comparison methods.

The five stimuli of any one context were not only judged individually, but in all possible combinations of two. Since there are ten possible pairs, there were fifteen stimuli in all. In the categorical assignments condition fifteen judgments were required. However, for the paired comparisons task, not all pairings among the fifteen stimuli were used. Following Thurstone and Jones (1959), those pairs in which both members contained a common element were omitted (e.g., a comparison between A vs. AB, or AB vs. AC). All other comparisons were utilized (e.g., A vs. B, A vs. BC, AB vs. CD). There were 55 pairs resulting.

Task order effects were controlled by distributing contexts and dimensions through a Latin-square design. Judgment booklets were assembled in the following three sequences: behavior-violations-gifts; violations-gifts-behavior; and gifts-behavior-violations. The three contexts of these dimensions were rotated similarly, each appearing equally often first, second, and third in order of presentation. For example, a complete booklet might consist of low-context behavior statements, high-context traffic violations, and neutral-context gift judgments. Each judge got one set of each context and one set of each type of social object, in a Latin-square arrangement. Nine such schedule types result. Paired comparison and categorical assignment items were identical for each S. That is, each S was redundantly exposed to each stimulus, but by different methods. Methods were alternated in sequence with half of the judges making paired comparisons followed by categorical assignments, and half giving categorical assignments followed by paired comparisons. The two methods of exposing each topic context were juxtaposed. This produced eighteen different booklet assemblies.

The stimuli for any given task were arranged in a random order, and three different random orders were employed. For the paired comparison tasks, R. T. Ross's (1934) method of optimizing orders for the presentation of pairs was used. The final product of all booklet and stimulus arrangement devices were 108 (3 x 6 x 2 x 3) types of judgment booklets, each calling for 210 judgments involving the three stimulus materials and two judgment methods.

Procedure. Judgment booklets were prearranged for distribution in a restricted random order. Judgment booklets contained self-administering instructions and were presented to groups ranging from 11-34. As a sample, here are the instructions for the categorical judgments of the behavioral stimuli:

The purpose of this schedule is to investigate ratings of statements about human behavior. There are five such statements which you are to rate individually and in various combinations with one another. In the case of both single and combined statements, assume that they represent behavior protocols of different adult individuals. Your task is to rate individuals on the basis of their protocols. At the left of the next page next to each of the protocols will appear a space where you are to write the number of the rating which you feel best depicts the individual represented. The scale provided below is to be used in making your ratings:

40	50	60
Normal behavior with no signs of eccentricity	Moderately abnormal behavior, with moderate amounts of eccentricity and bizarreness	Extremely abnormal behavior, with extreme degrees of bizarreness

If you think a statement or statements about an individual represents normal behavior, you should place a number in the neighborhood of "40" on the line to the left of the protocol. If you feel that the behavior is extremely bizarre, then a "60" would be more appropriate. Feel free to use numbers above, below and in between those indicated above. The above scale is intended only as a guide. Thus, if you have already used a "40" and come to a statement still more normal in your estimation, then you should use some lower number to accurately represent your judgment. Likewise, if you have already used a "60" and come to behavior which you consider still more bizarre, then you should use some still higher number. Please be sure to rate all the response statements and in the order presented. You may now turn the page and begin.

Paired comparisons instructions for Gifts read as follows:

Instructions. In this task you are to make paired comparisons of articles which students might receive as birthday gifts. There are 5 such gifts which you are to compare individually and in various combinations with one another. Assume that you do not already possess any of the articles. Consider them as gifts for your own personal use, which may not be sold. For each comparison, you are to check the article or articles which you would, in general, prefer to own.

Data Analysis

Analyses of the data involves two discrete steps. First is the determination of the scale values of the fifteen points by the ordinary means. These will be described first for all three methods. Following this, the mode of zero-point determination will be described.

Paired comparisons data. The Thurstone and Jones (1959) modification of Thurstone's (1927) law of comparative judgment to be used for composite as well as single stimuli was adopted in simplified form. A detailed report of this technique is given by Thurstone and Jones (1959, pp. 203-204). This procedure was simplified in the present investigation by assuming the same discriminal dispersion regardless of whether or not a paired comparison involved composite or single stimuli. A standard deviation of differences of 1.00 was used, allowing scale separations to be determined directly from unit normal deviates corresponding to the proportion of the time one stimulus was judged greater than another. (Alternative analyses are also explored, as will be discussed below.)

A special technique devised by Thurstone and Jones (1959) was required for estimating *z*-score values for empty cells resulting from the restriction that no stimulus be compared with itself. This method requires the averaging of scale separations for all possible comparisons according to the general formula:

$$(S_g-S_h) = \frac{1}{q} \sum (S_g-S_y) - (S_h-S_y)$$

where (S_g-S_h) refers to the scale distance between any two stimuli; $\sum (S_g-S_y) - (S_h-S_y)$ symbolizes the summation of *z*-score differences between any pair of stimuli; and q represents the number of columns in which entries appear for both stimuli g and h, a value which may change for different combinations of g and h. These computations adjust directly obtained *z*-score differences as well as filling empty cells. From the resulting table of *z*-score values, the scale values are determined according to the usual procedures (e.g., Guilford, 1954; Torgerson, 1958). In the present instances, the values were expressed in terms of the standard error of the difference rather than in discriminal dispersion units. The value of the lowest stimulus is used as the computational zero in this procedure.

Categorical judgments. For analysis by the method of categorical judgments, the responses were grouped into categories as follows: 39 and below, 40-44, 45-49, 50-54, 55-59, 60 and above. A table was prepared giving for each stimulus the percentage of judgments in that category or a lower one. These percentages were then turned into z values, and the differences between stimuli in terms of these z values averaged for the five lower categories (all percentages for 60 and above being 100%). This procedure is superior in obviating the necessity of computing the scale values for category boundaries and in avoiding interpolating. It can be called the method of multiple standards. It proves to be essentially identical to the more usual method of successive intervals, also used. These scale values use the lowest of the 15 stimuli as the computational zero, and are expressed in units of the standard error of the comparison.

Mean ratings. The third metric used treats the categorical judgment data as numerical ratings, averaging the two digit numbers to obtain the mean rating for each of the fifteen stimuli. These values are of course expressed in the units of the original judgment assignment, with 50 typically an average value, 40 low, 60 high, and with freedom to move outside of these limits.

Zero-point computation. Rational origins were obtained for all of the above analyses. The first step was to obtain the ten estimates of the additive constant, c, for each set of data by solving the equation,

$$S_{ij} - (S_i+S_j) = c$$

where S_{ij} refers to any composite stimulus whose component stimuli are S_i and S_j. The mean of each such set of estimates is then added to each scale value of the particular set of items from which it was derived. The resulting values are regarded as scale distances from the rational zero point pertaining to that particular set of data. A complete description of this procedure is given by Thurstone and Jones (1959).

It was suggested earlier that the acceptance of additivity depended upon some degree of agreement or clustering of zero-point determinations. These clusterings are graphically displayed in Figs. 1, 2, and 3. As a quantitative description of the degree of clustering, the ratio between the standard deviations of zero-point determinations and the standard deviations of corresponding composite stimuli has been computed. Such ratios, referred to subsequently as "additivity indices," were computed for each stimulus set. The greater the additivity index, the less the dispersion of zero points relative to the dispersion of composite stimuli.

Insert Figs. 1, 2, and 3 about here

The Thurstone and Jones (1959) method of describing additivity was a scatter-plot of scale values (after addition of the additive constant) with summated component values on the abscissa and corresponding composite values on the ordinate. That is, composite stimuli, S_{ij}, were plotted against their respective summated component stimuli, S_i+S_j. Presumably, the higher this correlation, the closer the fit to the additivity model. However, such an analysis was used only for comparative and supplementary purposes in this investigation. A high correlation is possible even when scale values of composite stimuli fall between or below those of their component stimuli, and the scattering of zero-point inferences is not directly examined.

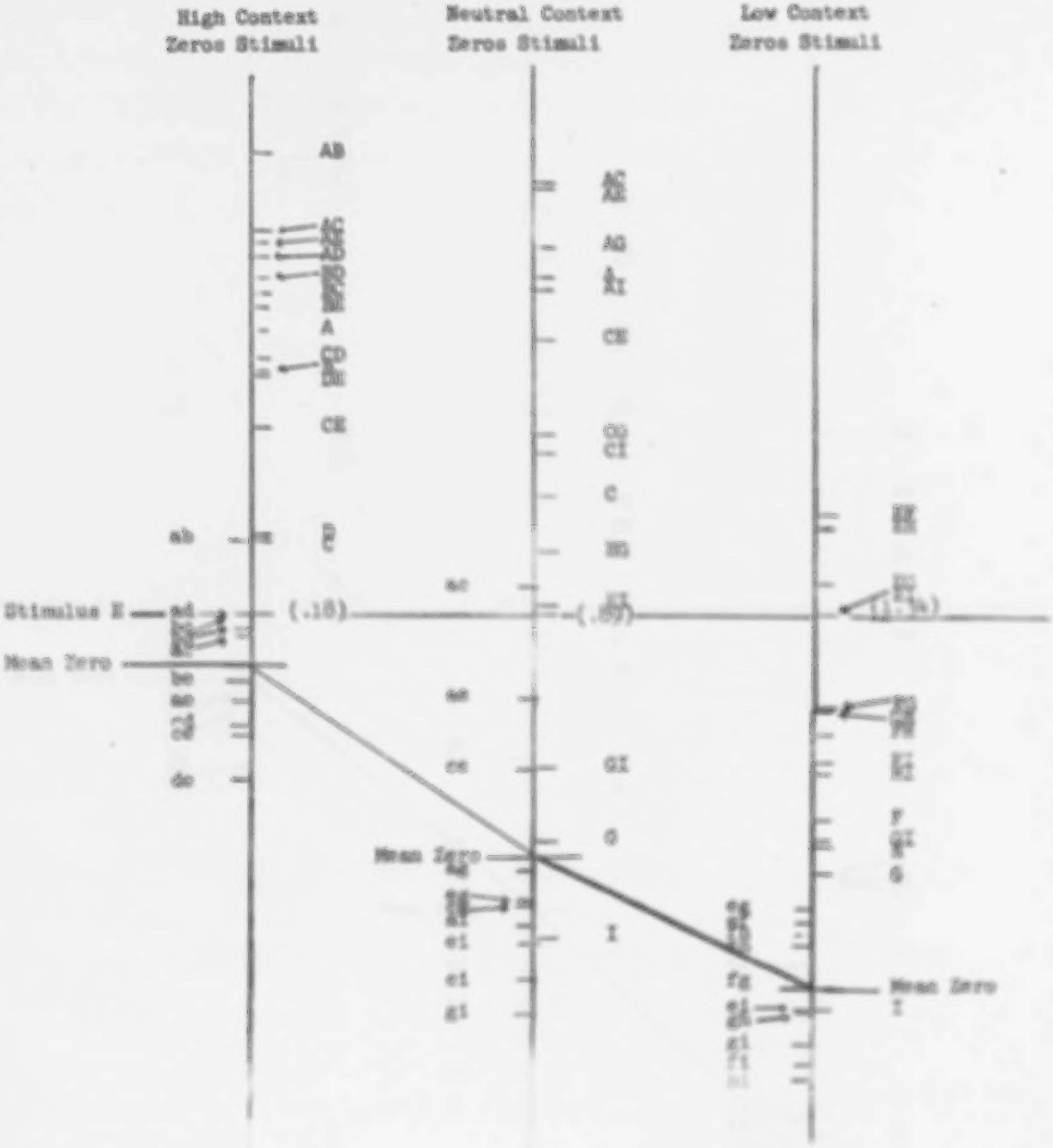


Fig. 1 Scale values and zero-point determinations of paired comparisons Behavior statements.

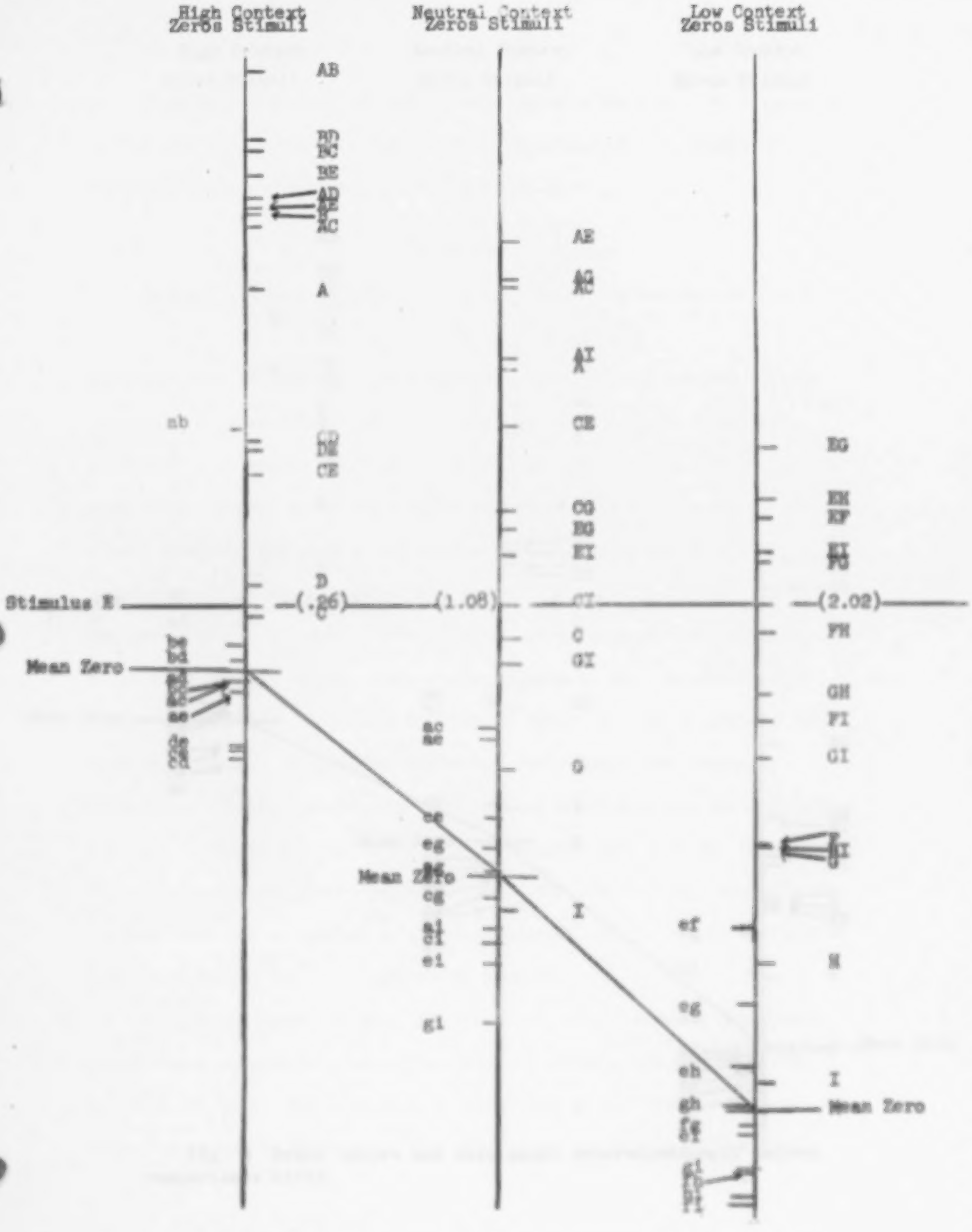


Fig. 2 Scale values and zero-point determinations on paired comparisons Traffic Violations.

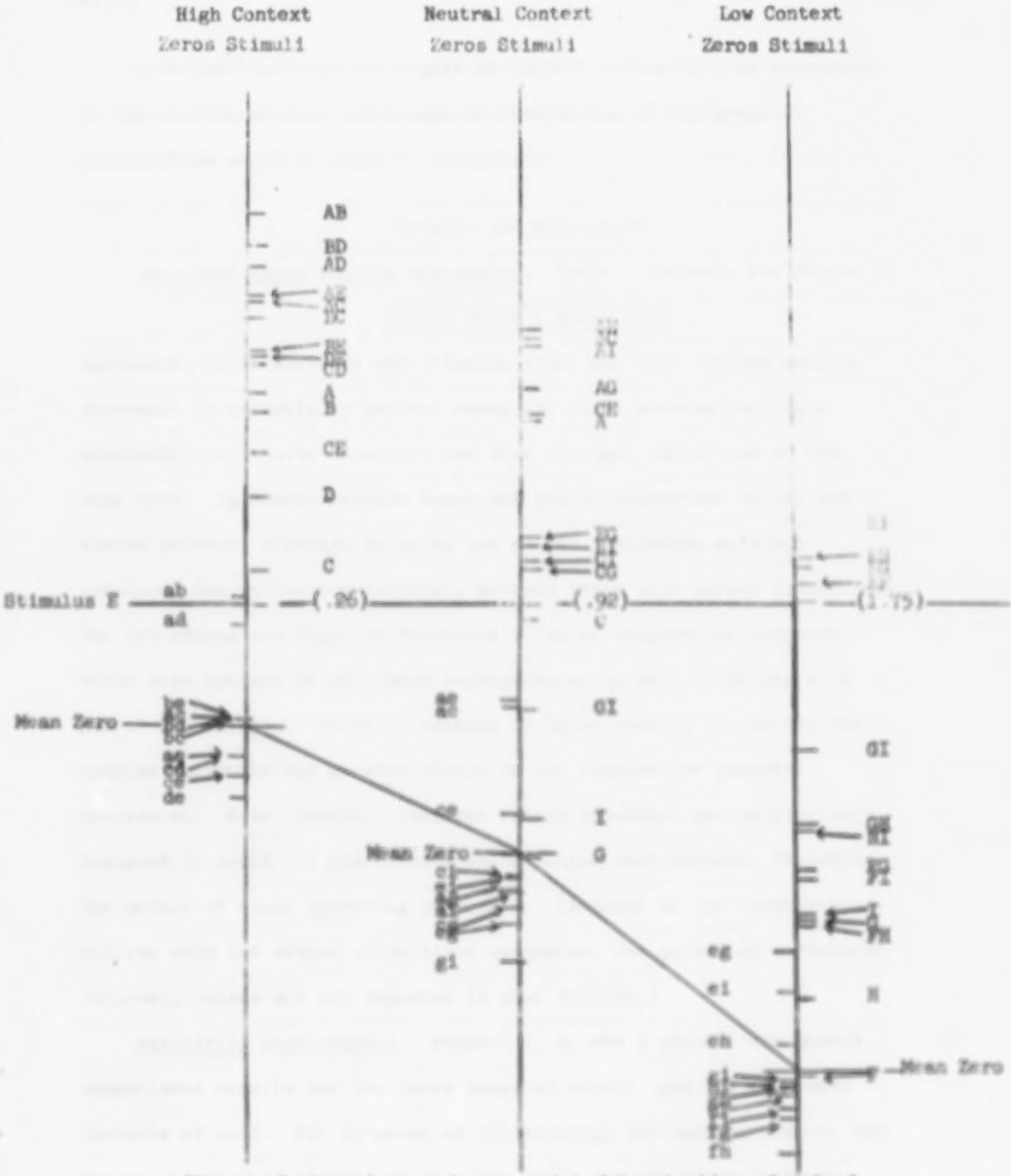


Fig. 3 Scale values and zero-point determinations of paired comparisons Gifts.

A further index of the degree of context effect will be presented in the results section subsequent to examination of the graphic presentation which it seeks to summarize.

Results and Discussion

Agreement among scaling procedures. Table 1 presents the degree of

Insert Table 1 about here

agreement within contexts and stimulus types for four scaling methods. Agreement is essentially perfect among the three methods (multiple standards, successive intervals and mean ratings) making use of the same data. Agreement between these and paired comparison values are almost perfect, although going as low as .96. Although multiple standards and successive intervals methods share with paired comparisons the procedures and logic of Thurstone's law of comparative judgment, while mean ratings do not, mean ratings agree as well or better with paired comparisons. There is nothing in these results to justify the complex rationale and greater effort of the comparative judgment procedures. Note, however, that the rating procedure was particularly designed to avoid the end effects which plague such methods, including the method of equal appearing intervals. (Because of the computational overlap with the method of multiple standards, the method of successive intervals values are not reported in what follows.)

Additivity requirements. Figures 1, 2, and 3 present the paired comparisons results for the three types of object and for the three contexts of each. For purposes of illustrating the context effect, the tracer stimulus is plotted at the same level for all three contexts, the other values being located in relation to it. Scale distances from the mean zero are indicated numerically for each E stimulus. Identification letters of stimuli are indicated in uppercase letters to the right of the vertical line. Zero-point locations are indicated to the left of the line in lowercase letters corresponding to the stimuli on which they are based. In addition to the three plotted, essentially identical results obtain for the multiple standards and the mean ratings computations. In terms of additivity requirements, we are thus discussing 27 zero-point studies (3 object types x 3 contexts x 3 scaling methods). In only five minor instances did a higher scale position obtain for a singlet than for a doublet of which it was a component. One of these is shown in Fig. 1, for the A stimulus, neutral context. In 25 of the 27 context set analyses, additivity indices exceeded unity as shown in Table 2. Assuming the probability of $\underline{SD}_D$ being less than $\underline{SD}_C$ on the basis of chance alone to be .50, the obtained .93 would be highly significant ($P < .001$), if one could assume independence. As predicted, Birthday Gifts yielded the highest combined additivity index of 1.98. Bizarre Behavior statements, expected to be least additive, showed an additivity index of 1.77, whereas the index for Traffic Violations was only 1.37. This surprise is traceable to categorical judgments data where a combined average Behavior index for multiple standards and mean ratings analyses was 1.94, and the combined average Violations index was a low 1.26. For the paired comparisons data, additivity indices ranked according to prediction with 1.82, 1.62, and 1.45 representing Gifts, Violations, and Behavior respectively. All

Insert Table 2 about here

but three stimulus sets fell short of an additivity index of 2.26 computed for the Thurstone-Jones (1959) data. The highest single additivity index produced by the present study was a 3.24 from low context multiple standards Behavior data. As expected, additivity indices for mean ratings correspond more closely to multiple standards than paired comparisons indices.

Note that the agreement among methods on additivity is not as high as for scale values within contexts. There is no uniform superiority for any one scaling method, multiple standards showing both the highest and lowest values.

Thurstone and Jones' assessment of the tenability of the additivity assumption was based upon scatter diagrams and correlation coefficients relating the composite stimulus with the sum of the two component stimuli. Such values have been computed for the present data. Figures 4 and 5 present the worst and the best of these. The correlations of .84 and .99 correspond to additivity indices of .95 and 2.80. The average correlation was .95.

Insert Figs. 4 and 5 about here

Context effects. Figures 1, 2, and 3 show striking evidence of context effects. These effects are equally striking when the data from multiple standards and mean ratings are graphed (not shown here). While the graphic displays are convincing, it is difficult to use them for making comparisons as to degree of context effect. For this reason, an index of context effect has been computed, in which the standard deviation of the three tracer stimulus E values has been divided by the average standard deviation within each of the three contexts of its fifteen stimuli. That is, the variability of level across contexts has been expressed

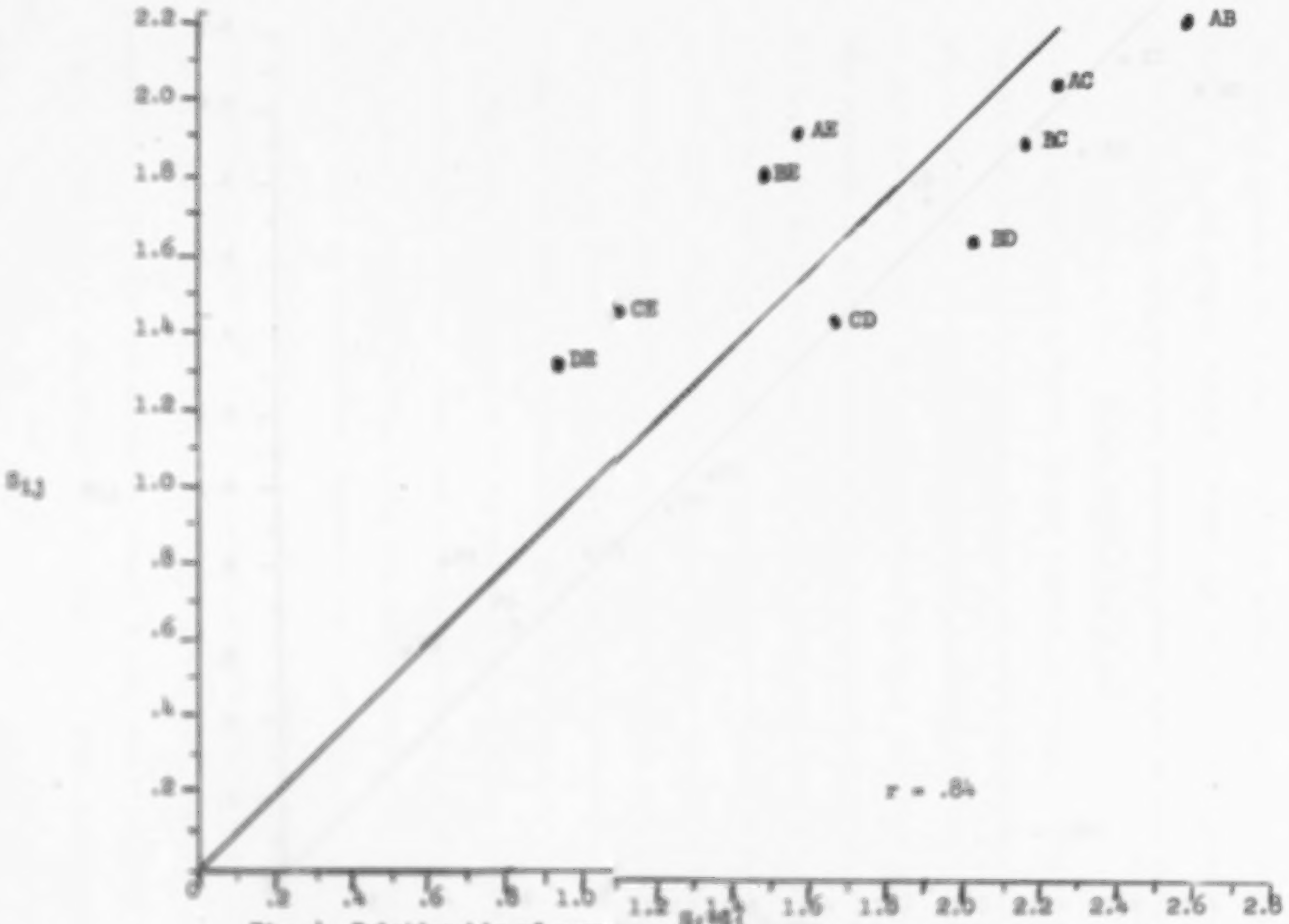


Fig. 4 Relationship of composite stimuli and summated components, high context multiple standards Behavior statements.

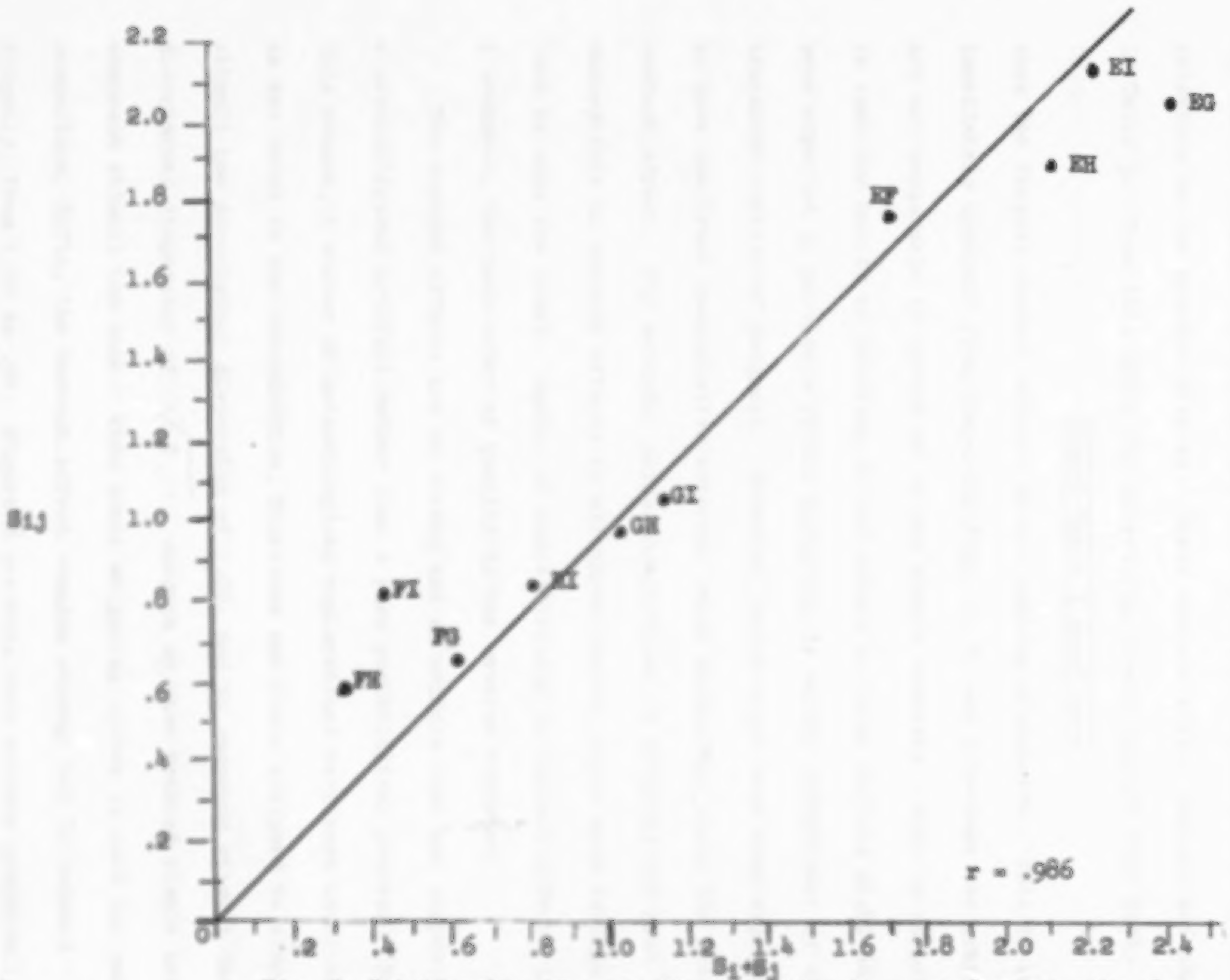


Fig. 5 Relationship of composite stimuli and summated components, low context multiple standards Gifts.

relative to the spread of values within contexts. A similar index using variances instead of standard deviations gave essentially similar results: the use of standard deviations was felt to better represent the spatial relations in the graphic display. These context effect indices are shown in Table 3. From this table the surprising result appears that Gifts

Insert Table 3 about here

show the largest context effect, by all scaling procedures. This is not immediately apparent from comparing Figs. 1, 2, and 3 because the graphs are not comparable in spread of values within contexts. This is surprising in that the habits of relating dollar values to these objects might have been expected to provide a stable anchoring of values independent of the transient context of judgment. Behavior, which might have been expected to have the least quantitative external value anchoring, shows the least context effect. For methods, paired comparisons is slightly the most susceptible to context effects in all three topics, while mean ratings tend to show the least. Again, if susceptibility to context effects is a weakness, the rank order of quality is the reverse expected.

The context effects are so strong and so uniform that one suspects a methodological artifact rather than a pure psychological process. For this reason, a number of methodological explorations have been undertaken. As was noted in the introduction, Thurstone and Jones assigned to simple stimuli the discriminal dispersion of 1.00, and to compound stimuli the discriminal dispersion of $\sqrt{2}$, whereas we have treated simple and compound stimuli the same. When their weighting system is used for paired comparison, Gifts, the context effect remains strong, but is reduced slightly, from 1.10 to .94. Figure 6 presents this outcome graphically.

The context effect is not judged to be appreciably diminished. The

Insert Fig. 6 about here

additivity index within contexts is, however, greatly increased: for High, Neutral, and Low contexts the values are, respectively 3.14, 2.26, and 3.47, averaging 3.29 compared to the average value of 1.82 for the uniform weighting procedure, shown in Table 2.

The Thurstone and Jones procedure assumes greater variability in appearance, greater discriminal dispersions, for compound stimuli, to an amount based on the formula for the standard deviation of a sum. If we assume that the variability in question is solely due to the fluctuating appearance to a single observer, it may seem reasonable that the variability of the composite be greater than that of a singlet. However, for the actual situation of application of the law of comparative judgment (for Thurstone's cases 2, 3, 4, and 5) individuals are pooled. Here the basic variability undoubtedly comes more from the stable differences across individuals in their stable preference hierarchies of the stimuli, than from the fluctuating appearance to any one individual. In this instance it probably is the case that stable individual differences in hierarchy of values are smaller for compounds of objects than for single objects, and hence discriminal dispersions smaller rather than greater. Whatever the case may be, a Thurstone case 3 analysis that infers stimulus differences in discriminal dispersions would be preferable to either a priori assumption. Such an analysis, following Burros (1951) as described by Guilford (1954) has been done for paired comparisons, Gifts. For this, the index of context effect is 1.05, but a slight improvement on the 1.10 value previously found. Because the picture remains essentially identical, a graphic

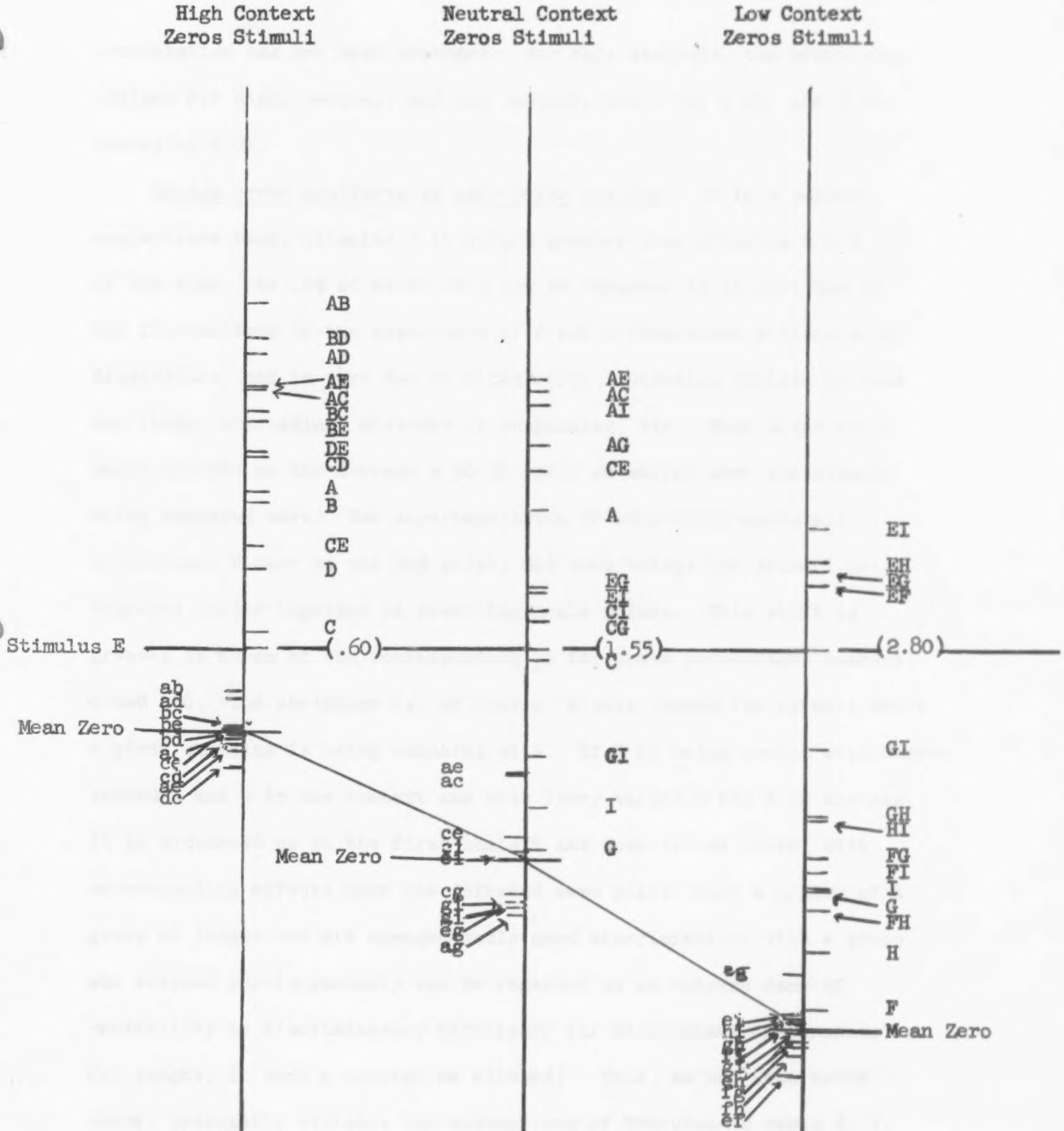


Fig. 6 Thurstone-Jones Weighted scale values and zero-points of paired comparisons Gifts.

presentation has not been provided. For this analysis, the additivity indices for high, neutral, and low contexts are 1.72, 1.63, and 2.78, averaging 2.01.

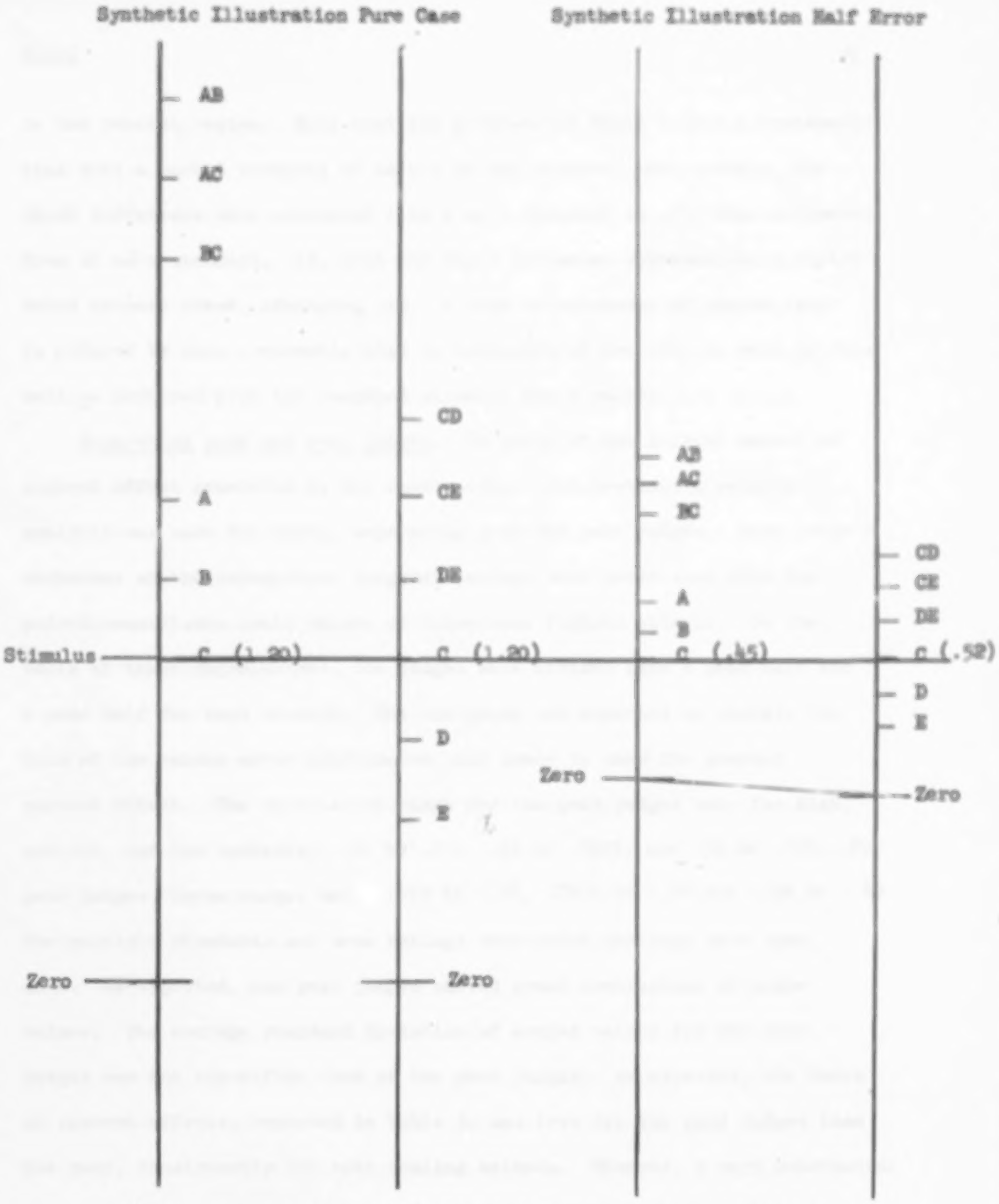
Random error artifacts in zero-point scaling. If in a paired-comparisons task, stimulus A is judged greater than stimulus B 90% of the time, the 10% of exceptions can be regarded as in part due to the fluctuations in the appearance of A and B (Thurstone's discriminal dispersions) and in part due to blind error, including failure to read the items, misreading, mistakes in responding, etc. Such blind error would provide on the average a 50-50 split no matter what the stimuli being compared were. The superimposition of such error moves all comparisons closer to the 50% point, and thus brings the stimuli being compared closer together in resulting scale values. This shift is greater in terms of the corresponding zs for those percentages nearest 0 and 100. The shrinkage is, of course, always toward the stimuli which a given stimulus is being compared with. If C is being scaled with higher valued A and B in one context and with lower valued D and E in another, it is distorted up in the first context and down in the other, with corresponding effects upon the inferred zero point. Such a mixing of a group of judges who are homogeneously good discriminators with a group who respond purely randomly can be regarded as an extreme case of variability in discriminatory efficiency (or discriminal dispersions for judges, if such a concept be allowed). This, as has been noted above, presumably violates the assumptions of Thurstone's cases 2, 3, 4, and 5.

Because the problem is complex and because the magnitude of the effect cannot be anticipated easily, a specific synthetic example is worked out in Fig. 7 and Tables 4, 5, and 7. This starts from a set of pure absolute scale values (stated in units of the standard error

Insert Fig. 7 and Tables 4, 5, 6, and 7 about here

of differences) of: A = 1.80, B = 1.50, C = 1.20, D = .90, and E = .60. These, and the corresponding composite stimulus values are graphed on the left of Fig. 7. A, B, and C are scaled together in the high context, C, D, and E for the low. This a priori model generates the "pure" _z_ values for comparisons of Table 4, which in turn are transformed into the proportions of judgments of greater than of Table 5. Now consider a population of judges half of whom perform as in Table 5, the other half of whom judge blindly, whose table corresponding to Table 5 would have all entries at .500. The proportions for the pooled population would be as shown in Table 6, generating the _z_s of Table 7, with the resultant scale values shown there. These are graphically displayed on the right side of Fig. 7. As can be seen, a great shrinkage of scale occurs, and a slight context effect. The index of context effect here is .07, which corresponds to that for paired comparisons, Gifts, of 1.42, when the latter is correspondingly computed for high and low contexts only.

The illustration utilized an unreasonably large proportion of random error. The degree of context effect generated is too small to account for the results obtained in our empirical cases. The synthetic illustration does indicate, however, that the problem is technically relevant for all uses of the law of comparative judgments. Note that the gaps between extreme stimuli are more shrunk than between stimuli



Synthetic Illustration Pure Case

Synthetic Illustration Half Error

AB
AC
BC
A
B
Stimulus
C (1.20)
Zero
CD
CE
DE
C (1.20)
D
E
Zero
AB
AC
BC
A
B
C (.45)
Zero
CD
CE
DE
C (.52)
D
E
Zero

Fig. 7 Synthetic Illustration of artifactual context effect from Pure Case to Half Error.

in the central region. Note that the z values of Table 7 show a systematic bias with a marked cramping of values in the corners. For example, the AB-AC difference when estimated from C as a standard is .03, when estimated from AC as a standard, .15, with the other estimates systematically distributed between these, averaging .10. A clue to existence of random error is offered by this systematic bias in estimates of the cell by cell zs from cell zs inferred from the computed stimulus scale values.

Separating good and poor judges. In spite of the trivial amount of context effect generated by the random error illustration, a related analysis was made for Gifts, separating good and poor judges. Each judge's responses on the categorical judgment ratings were correlated with the paired-comparisons scale values of these same fifteen stimuli. On the basis of these correlations, the judges were divided into a good half and a poor half for each context. The low group was expected to contain the bulk of the random error performance, and hence to show the greater context effect. The correlation range for the good judges was, for high, neutral, and low contexts, .94 to .575, .99 to .7975, and .99 to .753. For poor judges, these ranges were: .574 to -.48, .7974 to -.66 and .752 to -.40. The multiple standards and mean ratings zero-point scalings were then done. As expected, the poor judges showed great contraction of scale values. The average standard deviation of scaled values for the poor judges was but two-fifths that of the good judges. As expected, the index of context effects, reported in Table 3, was less for the good judges than the poor, consistently for both scaling methods. However, a very substantial context effect remains for the good judges, comparable to that shown in Figs. 6 or 2. Thus data quality does not seem adequate to explain the basic context effect findings.

Context effects in terms of response-scale values. The usual statement of context effects is in terms of the judges' response language, whether stated as the value of a point of subjective equality or as the mean judgment given to a tracer stimulus. It should both provide continuity with that tradition, and help clarify the psychological meaningfulness of the context effects found, to present our findings in these terms. Figures 8, 9, and 10 present these results. The most direct representation of the context effect is in the line showing the locations of the tracer stimulus E. In all three graphs, this shows the higher value for the lower context, the typical result.

Insert Figs. 8, 9, and 10 about here

Were the context effect to have been complete, then the mean stimulus value within each context would have been the same. This has been found to be the case in some studies (e.g., Krantz & Campbell, 1961). Here, however, the mean ratings reflect in part and in rank order the differences in the stimuli being scaled. This may be due to the external referents in the definition of the response language terms. For example, the value 50 was defined for Behavior "Moderately abnormal behavior, with moderate amounts of eccentricity and bizarreness." For Traffic Offenses, 50 was "A moderately serious offense." For Gifts, 50 was defined as "Quite desirable, would enjoy receiving as a gift." That the independence from context was only partial, is shown by the shifting values for the tracer E. The phenomenal reality of these differences is enhanced by the external referents of the response definitions (e.g., Campbell, Lewis, & Hunt, 1958).

The values for the computed zero points when stated in the original response language follow closely the trends of the mean stimulus value,

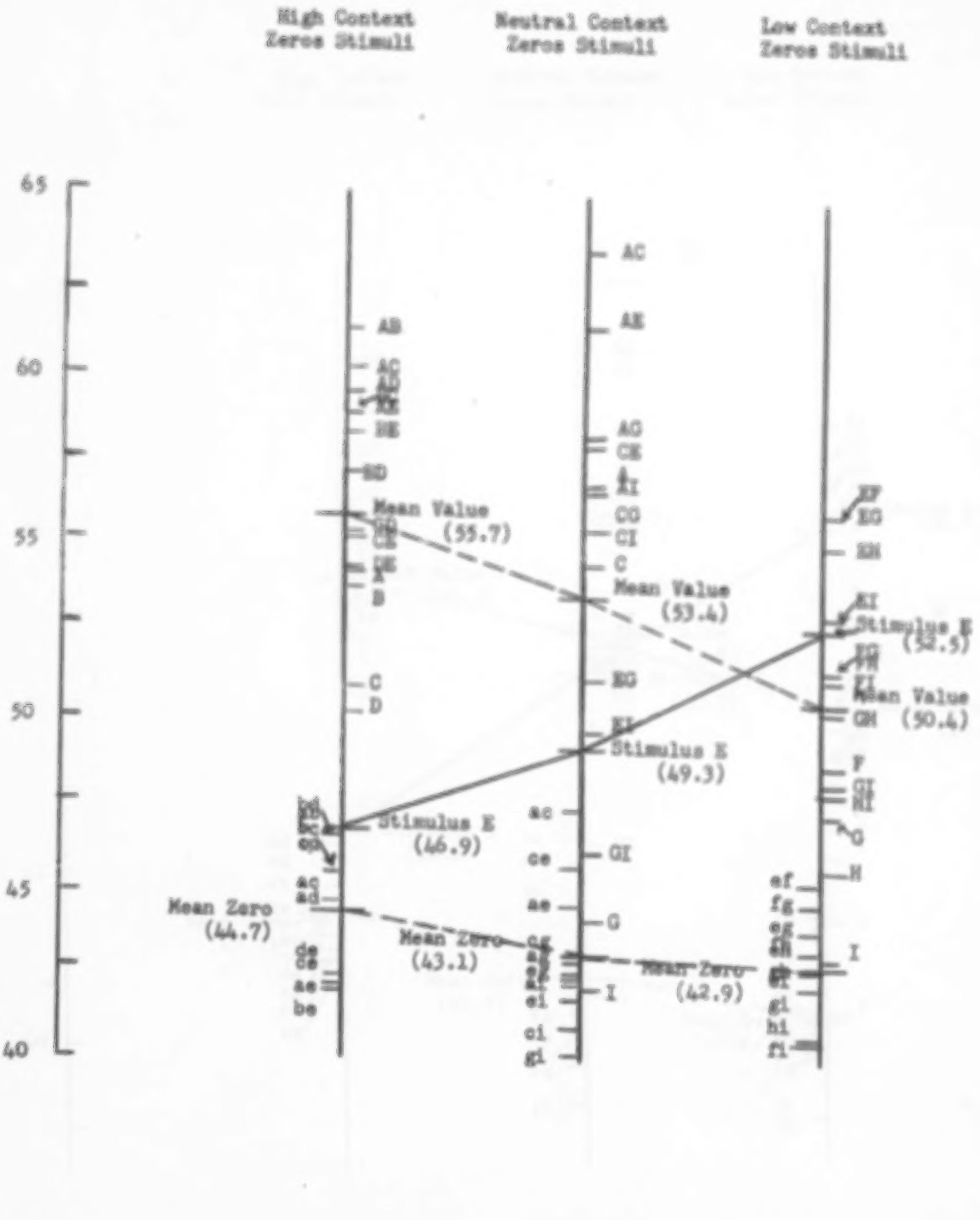


Fig. 8 Original Judgmental Values for Behavior statements

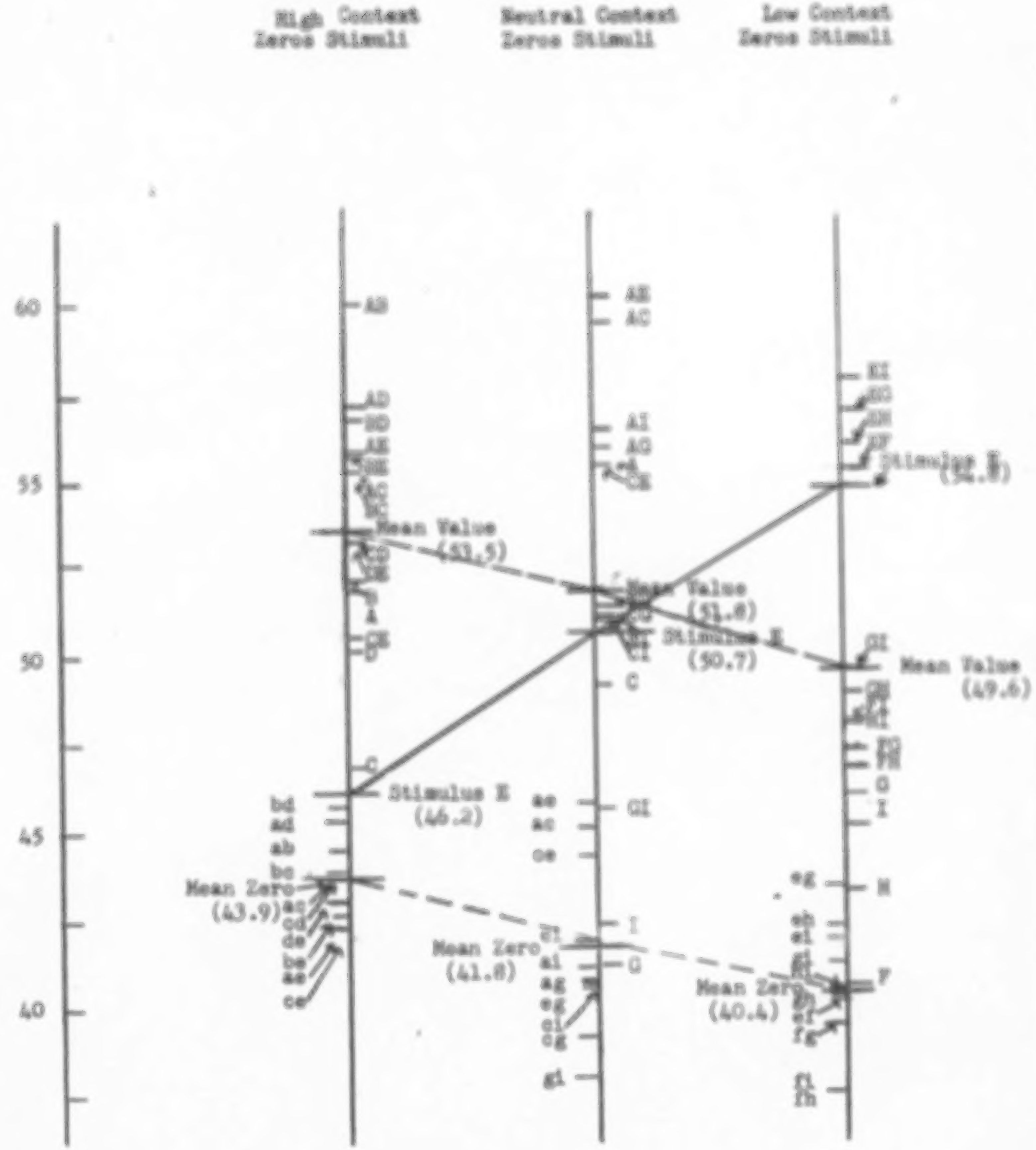


Fig. 9 Original Judgmental Values for Gifts

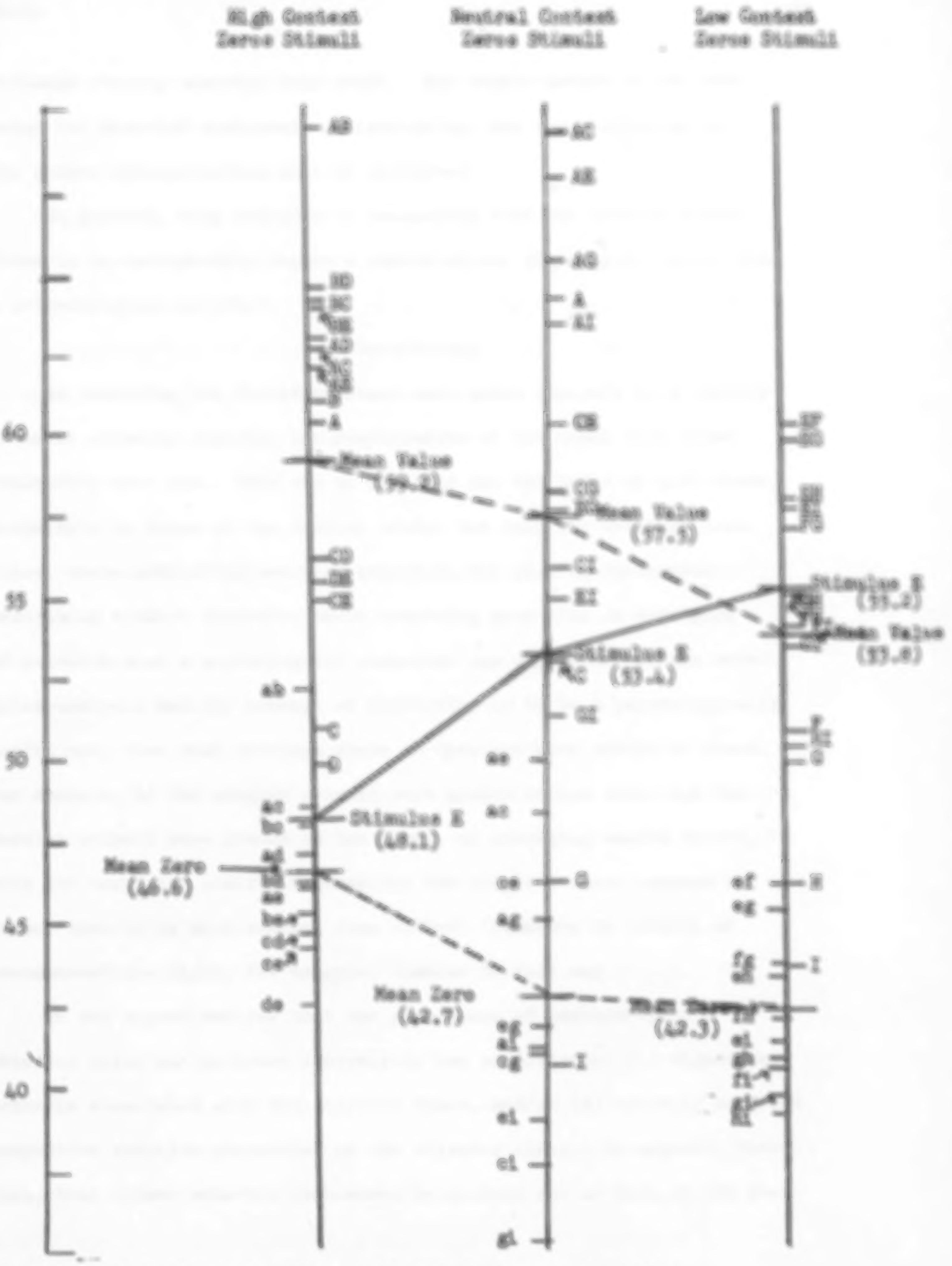


Fig. 10 Original Judgmental Values for Violations

although showing somewhat less shift. The stable nature of the zero point for Behavior statements is intriguing, but speculation as to its proper interpretation will be postponed.

In general, this analysis is reassuring that the context effect found is in considerable degree a psychological phenomenon, rather than a methodological artifact.

Conclusions

In extending the Thurstone-Jones zero-point analysis to a variety of novel stimulus objects, the requirements of the model were found reasonably well met. This was so not only for the topic of gift items, comparable to those of the initial study, not only for traffic violations, where additivity was also expected, but also to statements describing bizarre behavior, where something more like an averaging of evidence than a psychological summation was expected. If the zero-point analysis and the concept of additivity is to be a psychologically useful one, then some settings where it does not work should be found. For example, if the singlet stimuli were grades on one exam, and the doublet stimuli were grades on two exams, an averaging should result, with the composite scaling between the two singlets which compose it rather than being more extreme than either. Comments in letters of recommendations might, for example, combine in this way.

It was argued earlier that the phenomenon of additivity of a stimulus class was governed entirely by two conditions: (1) objective criteria associated with the stimulus class, and/or (2) socially anchored subjective additive properties of the stimulus class. It appears, therefore, that either behavior statements do contain one or both of the above attributes, or a third variable, must be considered; viz., the capacity to transfer learned additive skills to novel stimuli. In this latter case, additivity could be achieved in the absence of both of the above conditions, or in conjunction with one or both of them. A corollary possibility may be an interaction between additivity transfer and the presence of demand characteristics of the experiment suggested by judges' instructions and nature of task presentation. That is, a judge may add or average, depending upon possible desired outcomes implied to him by the experimental arrangement. Our instructions, and the joint presence of other tasks in which additivity was appropriate, may have worked in this direction.

A very desirable feature in a meaningful psychological zero point would be stability, so that the same zero point would be inferred from different sets of stimuli. Indeed, the zero point would hardly seem meaningful without showing invariance over some multiple methods of ascertainment. On the other hand, if psychological valuations do genuinely change with context, then the zero point should perhaps change accordingly. Whatever the interpretation, the zero points in the present study were markedly affected by context. While a trivial amount of this effect may be attributed to a random-error artifact, in the main, it seems a psychological phenomenon.

Several methods of scaling were employed, all producing equivalent results. A simple mean rating procedure provided excellent zero-point determinations, and correlated .96 or better with other methods. This procedure was likewise as effective as the others in maximizing additivity and minimizing context effects. The methods of successive intervals and multiple standards were essentially equivalent to each other and to paired comparisons.

References

Burros, R. H. The application of the method of paired comparisons to the study of reaction potential. *Psychol. Rev.*, 1951, 58, 60-66. Cited in J. P. Guilford, *Psychometric Methods* (2nd ed.) New York: McGraw-Hill, 1954.

Campbell, D. T , Lewis, N., & Hunt, W.A. Context effects with judgmental language that is absolute, extensive, and extra-experimentally anchored. *J. exp. Psychol.*, 1958, 55, 220-228.

Guilford, J. P. *Psychometric methods* (2nd ed.) New York: McGraw-Hill, 1954.

Helson, H. Adaptation level as frame of reference for prediction of psychophysical data. *Amer. J. Psychol.*, 1947, 60, 1-29.

Helson, H. Adaptation level as a basis for a quantitative theory of frame of reference. *Psychol. Rev.*, 1948, 55, 297-313.

Helson, H. Adaptation-level theory. In S. Koch (Ed.), *Psychology: a study of a science*. Vol. 1. New York: McGraw-Hill, 1959. Pp. 565-621.

Jones, L. V. A rational origin in paired comparison and successive interval scaling. Presented at the Symposium, "Recent developments in psychometric scaling." New York, Sept. 4, 1957.

Krantz, D. L., & Campbell, D. T. Separating perceptual and linguistic effects of context shifts absolute judgments. *J. exp. Psychol.*, 1961, 62, 35-42.

Osgood, C. E., & Sebeok, T. A. (Eds.) Psycholinguistics: a survey of theory and research problems. Supplement to *J. abnorm. soc. Psychol.*, 1954, 49, No. 4, Part 2, pp. 1-203.

Rimoldi, H. J. A. Prediction of scale values for combined stimuli. Brit. J. stat. Psychol., May 1956. Cited in L. V. Jones, A rational origin in paired comparison and successive interval scaling. Presented at the Symposium, "Recent developments in psychometric scaling." New York, Sept. 4, 1957.

Ross, R. T. Optimum orders for the presentation of pairs in the method of paired comparisons. J. educ. Psychol., 1934, 25, 375-382.

Shuford, E. H., Jones, L. V., & Bock, R. D. A rational origin obtained by the method of contingent paired comparisons. Psychometrika, 1960, 343-356.

Thurstone, L. L. A law of comparative judgment. Psychol. Rev., 1927, 273-286.

Thurstone, L. L., & Jones, L. V. Measuring subjective values. In L. L. Thurstone, The measurement of values. Chicago: Univer. of Chicago Press, 1959.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

16

Table 1

Scaling Method Intercorrelations

	Multiple Standards	Successive Intervals	Raw Mean Ratings
High context behavior			
paired comparisons	.964	.962	.961
multiple standards		.999	.997
successive intervals			.998
Neutral context behavior			
paired comparisons	.983	.982	.980
multiple standards		.999	.999
successive intervals			.999
Low context behavior			
paired comparisons	.961	.961	.963
multiple standards		.998	.995
successive intervals			.998
High context violations			
paired comparisons	.979	.980	.986
multiple standards		.999	.999
successive intervals			.999
Neutral context violations			
paired comparisons	.975	.972	.986
multiple standards		.999	.997
successive intervals			.995
Low context violations			
paired comparisons	.978	.975	.983
multiple standards		.999	.997
successive intervals			.994
High context gifts			
paired comparisons	.981	.983	.983
multiple standards		.999	.996
successive intervals			.994
Neutral context gifts			
paired comparisons	.996	.995	.989
multiple standards		.999	.996
successive intervals			.996
Low context gifts			
paired comparisons	.990	.991	.992
multiple standards		.999	.998
successive intervals			.998

Table 2

Additivity Indices

Dimension	Context	Paired Comparisons SD_c	 SD_o	 Index	Multiple Standards SD_c	 SD_o	 Index	Mean Ratings SD_c	 SD_o	 Index	Combined Average Additivity Indices
Behavior	high	.272	.241	1.13	.286	.302	.95	2.22	1.69	1.31	
Behavior	neutral	.662	.459	1.44	.770	.348	2.21	5.05	2.06	2.45	
Behavior	low	.380	.209	1.82	.424	.131	3.24	2.77	1.41	1.96	
Means		.438	.303	1.45	.493	.260	1.90	3.35	1.72	1.95	1.77
Violations	high	.552	.352	1.57	.532	.356	1.49	4.44	2.62	1.69	
Violations	neutral	.653	.360	1.81	.557	.610	.91	5.35	4.46	1.20	
Violations	low	.491	.333	1.47	.388	.366	1.06	2.90	2.16	1.34	
Means		.565	.348	1.62	.492	.444	1.11	4.23	3.08	1.37	1.37
Gifts	high	.245	.156	1.57	.288	.153	1.88	2.38	1.11	2.14	
Gifts	neutral	.456	.317	1.44	.529	.309	1.71	4.24	2.44	1.74	
Gifts	low	.549	.215	2.55	.568	.203	2.80	4.33	1.84	2.35	
Means		.417	.229	1.82	.462	.222	2.08	3.65	1.80	2.03	1.98
Grand Means		.473	.294	1.64	.480	.306	1.70	3.74	2.21	1.72	1.70
Thurstone-Jones Gifts		.497	.220	2.26	(unweighted)						

Table 3

Indices of Context Effect

	Behavior	Violation	Gifts
Paired Comparisons	.81	.96	1.10
Multiple Standards	.71	.90	1.08
Mean Ratings	.61	.79	1.09

	Unweighted	Weighted	Case III
Paired Comparisons Gifts	1.10	.94	1.05

	Good Judges	Poor Judges
Multiple Standards Gifts	.97	1.21
Mean Ratings Gifts	.88	1.29

Table 4

"Pure" z Values for Synthetic Illustration

	AB	AC	BC	A	B	C
AB	.00	.30	.60	1.50	1.80	2.10
AC	-.30	.00	.30	1.20	1.50	1.80
BC	-.60	-.30	.00	.90	1.20	1.50
A	-1.50	-1.20	-.90	.00	.30	.60
B	-1.80	-1.50	-1.20	-.30	.00	.30
C	-2.10	-1.80	-1.50	-.60	-.30	.00

	CD	CE	DE	C	D	E
CD	.00	.30	.60	.90	1.20	1.50
CE	-.30	.00	.30	.60	.90	1.20
DE	-.60	-.30	.00	.30	.60	.90
C	-.90	-.60	-.30	.00	.30	.60
D	-1.20	-.90	-.60	-.30	.00	.30
E	-1.50	-1.20	-.90	-.60	-.30	.00

Table 5

"Pure" Proportions for Synthetic Illustration

	AB	AC	BC	A	B	C
AB	.500	.618	.726	.933	.964	.982
AC	.382	.500	.618	.885	.933	.964
BC	.274	.382	.500	.816	.885	.933
A	.067	.115	.184	.500	.618	.726
B	.036	.067	.115	.382	.500	.618
C	.018	.036	.067	.274	.382	.500

	CD	CE	DE	C	D	E
CD	.500	.618	.726	.816	.885	.933
CE	.382	.500	.618	.726	.816	.885
DE	.274	.382	.500	.618	.726	.816
C	.184	.274	.382	.500	.618	.726
D	.115	.184	.274	.382	.500	.618
E	.067	.115	.184	.274	.382	.500

Table 6

"Half-Error" Proportions for Synthetic Illustration

	AB	AC	BC	A	B	C
AB	.500	.559	.613	.716	.732	.741
AC	.441	.500	.559	.692	.716	.732
BC	.387	.441	.500	.658	.692	.716
A	.284	.308	.342	.500	.559	.613
B	.268	.284	.308	.441	.500	.559
C	.259	.268	.284	.387	.441	.500

	CD	CE	DE	C	D	E
CD	.500	.559	.613	.658	.692	.716
CE	.441	.500	.559	.613	.658	.692
DE	.387	.441	.500	.559	.613	.658
C	.342	.387	.441	.500	.559	.613
D	.308	.342	.387	.441	.500	.559
E	.284	.308	.342	.387	.441	.500

Table 7

"Half-Error" Proportions for Synthetic Illustration

	AB	AC	BC	A	B	C	Σ	M	c	Absolute Scale Value
AB	.00	.15	.29	.57	.62	.65	2.28	.38	.83	1.21
AC	-.15	.00	.15	.50	.57	.62	1.69	.28	.83	1.11
BC	-.29	-.15	.00	.41	.50	.57	1.04	.17	.83	1.00
A	-.57	-.50	-.41	.00	.15	.29	-1.04	-.17		.66
B	-.62	-.57	-.50	-.15	.00	.15	-1.69	-.28		.55
C	-.65	-.62	-.57	-.29	-.15	.00	-2.28	-.38		.45

	CD	CE	DE	C	D	E	Σ	M	c	Absolute Scale Value
CD	.00	.15	.29	.41	.50	.57	1.92	.32	.59	.91
CE	-.15	.00	.15	.29	.41	.50	1.20	.20	.59	.79
DE	-.29	-.15	.00	.15	.29	.41	.41	.07	.59	.66
C	-.41	-.29	-.15	.00	.15	.29	-.41	-.07		.52
D	-.50	-.41	-.29	-.15	.00	.15	-1.20	-.20		.39
E	-.57	-.50	-.41	-.29	-.15	.00	-1.92	-.32		.27

Abstract

Four hundred eighty-two college students served as judges in the exploration of a method of determining a rational origin on an interval scale developed by Thurstone and Jones. This method is based on Thurstone's law of comparative judgment. It not only requires the judgment of single stimuli, but pairs as well. A basic assumption is that the scale values of pairs will equal the summation of scale values of corresponding components. The stimulus dimensions of birthday gifts, traffic violations, and bizarre behavior symptoms were studied. Only gifts and violations were predicted to be additive. Each dimension was presented in high, neutral, and low context sets. Paired comparisons and categorical assignments methods were employed. Each judge made 210 judgments involving instances of all three contexts, each represented by a different dimension.

Support for the zero-point model was found in all three dimensions. In line with prediction, gifts were most satisfactorily additive. Evidence of additivity was found for behavior symptoms, contrary to prediction. Strong context effects were found under all conditions and modes of analysis. While a small proportion of this context effect could be interpreted as a random error artifact, in bulk it seems a psychological effect. The simplest of four modes of analysis used, mean ratings, was as effective as any of the more complex methods in maximizing additivity and minimizing context effects.

Appendix

for

Zero-Point Scaling as Affected by Social Object, Scaling Method, and Context

Table of Contents

Identification of Topic Stimuli 1

Behavior, Violations,Gifts-Ratio Scale Values and O-Point Deviations from Mean O-Point for Means, Paired Comparisons, and Multiple Standards. 2

Gifts (Weighted) - Ratio Scale Values and O-Point Deviations from Mean O-Point for Paired Comparisons. 5

Gifts (Case III) - Ratio Scale Values and O-Point Deviations from Mean O-Point for Paired Comparisons 6

Gifts (Good and Poor Judges) - Ratio Scale Values and O-Point Deviations from Mean O-Point for Raw Means and Multiple Standards. 7

Behavior, Violations, Gifts-Proportion Matrices for Paired Comparisons and Multiple Standards 9

Gifts (Good and Poor Judges) - Proportion Matrices for Multiple Standards 27

Topic Stimuli

Gifts

A	TV Set
B	Record Player
C	Camera
D	Typewriter
E	Transistor Radio
F	Leather Binder
G	Desk Lamp
H	Pen and Pencil Set
I	Dictionary

Violations

A	Driving under the Influence of Alcohol
B	Leaving the Scene of an Accident
C	Driving after License has been Revoked
D	Driving 50 mph. in a 25 Mile Zone
E	Running a Red Light
F	Running a Stop Sign
G	Going the Wrong Way on a One-Way Street
H	Following a Car too Closely
I	Failing to Signal Before Turning

Behavior

A	He tried to squeeze his head between the bars of his bed.
B	He kept hiding in dark corners and crying.
C	He said the green walls of his room were really purple.
D	He went on a number of hunger strikes.
E	He often said, "I wish I were dead."
F	One day he brushed his teeth six times.
G	He seldom smiled.
H	He was once found walking in his sleep.
I	He often complained of a backache.

BEHAVIOR

Ratio Scale Values and 0-point Deviations
from mean 0-point

Scale Values

Means			Paired Comparisons			Multiple Standards		
H	N	L	H	N	L	H	N	L
AB 16.9	AC 20.4	EF 12.9	AB 1.85	AC 2.44	EF 1.71	AB 2.23	AC 3.10	EF 1.85
AC 15.9	AE 18.1	EG 12.9	AC 1.57	AE 2.42	EH 1.67	AC 2.06	AE 2.86	EG 1.83
AD 15.1	AG 15.1	EH 11.9	AE 1.53	AG 2.20	EG 1.46	AD 1.98	AG 2.32	EH 1.65
AE 14.5	CE 14.7	EI 9.7	AD 1.49	A 2.10	EI 1.34	AE 1.93	CE 2.31	EI 1.34
BC 14.5	A 13.5	E 9.5	BD 1.40	AI 2.06	E 1.34	BC 1.92	AI 2.13	E 1.34
BE 14.0	AI 13.3	FG 8.4	BC 1.35	CE 1.88	FG 1.01	BE 1.80	A 2.08	FG 1.19
BD 12.7	CG 12.3	FI 8.1	BE 1.30	CG 1.53	GH 1.00	BD 1.65	CG 1.87	FI 1.11
CD 11.0	CI 12.3	FH 8.1	A 1.22	CI 1.46	FH .92	CE 1.44	CI 1.86	FH 1.09
CE 10.8	C 11.1	GH 7.1	CD 1.12	C 1.31	FI .82	CD 1.43	C 1.73	GH .96
DE 9.9	EG 7.8	F 5.7	B 1.07	EG 1.10	HI .77	A 1.34	EG 1.24	F .78
A 9.8	EI 6.4	GI 5.1	DE 1.04	EI .91	F .61	DE 1.33	EI 1.00	GI .63
B 9.3	E 6.2	HI 4.9	CE .87	E .89	GI .54	B 1.27	E .97	HI .60
C 6.5	GI 3.0	G 4.3	D .46	GI .33	H .51	C .88	GI .51	G .49
D 5.7	G 1.1	H 2.7	C .45	G .06	G .42	D .74	G .24	H .36
E 2.5	I -1.0	I 0.0	E .18	I -.28	I -.06	E .21	I -.22	I .09
Var.								
15.80	35.80	12.89	.20	.69	.23	.29	.85	.26
S.D.								
3.97	5.98	3.59	.45	.83	.48	.54	.92	.51

Zero Points

H	N	L	H	N	L	H	N	L
bd 2.27	ac 4.22	ef 2.35	ab .445	ac .977	eg .297	ab .384	ac .716	ef .272
ab 2.21	ce 2.53	fg 1.66	ad .183	ae .576	ef .235	bd .362	ce .388	ei .091
bc 1.22	ae 1.54	eg .91	bc .180	ce .323	fh .195	bc .236	ae .194	fg .082
cd 1.21	cg -.12	fh .29	bd .129	ag-.036	eh .174	cd .191	cg .092	fh .054
ac .42	ag -.50	eh .27	ac .109	eg-.147	fg .021	ac .154	ag-.002	eh .052
ad .42	eg -.61	gh -.06	be-.055	cg-.160	ei-.064	ad .094	eg-.033	eg-.006
de-1.65	ai -.80	ei -.20	ae-.130	ai-.240	gh-.074	be-.312	ei-.252	gi-.053
ce-1.78	ei-1.21	gi -.74	cd-.216	ei-.306	gi-.139	ce-.349	ai-.266	gh-.114
ae-2.16	ci-2.14	hi-2.14	ce-.244	ci-.436	fi-.274	ae-.379	ci-.347	hi-.143
be-2.1	gi-2.92	fi-2.34	de-.405	gi-.556	hi-.321	de-.381	gi-.495	fi-.235

Additivity Index

	SDc	SDo	Index	SDc	SDo	Index	SDc	SDo	Index
H	2.22	1.69	1.31	.272	.241	1.13	.286	.302	.95
N	5.05	2.06	2.45	.662	.459	1.44	.770	.348	2.21
L	2.77	1.41	1.96	.380	.209	1.82	.424	.131	3.24
Means	3.35	1.72	1.95	.438	.303	1.45	.493	.260	1.90

Combined Average Additivity Index 1.77

Context Effect Index

Var.	.35	.62	.47
S.D.	.61	.81	.71

VIOLATIONS

Ratio Scale Values and O-point Deviations from mean O-point

Scale Values

	Means			Paired Comparisons			Multiple Standards		
	H	N	L	H	N	L	H	N	L
	AB 22.8	AC 25.6	EF 18.0	AB 2.40	AE 2.74	EG 2.65	AB 2.92	AC 2.93	EF 2.47
	BD 18.2	AE 25.1	EG 17.6	BD 2.13	AG 2.59	EH 2.44	BD 2.35	AE 2.74	EG 2.42
	BC 17.9	AG 22.7	EH 15.8	BC 2.08	AC 2.56	EF 2.37	BC 2.24	AG 2.54	EH 2.17
	BE 17.6	A 21.7	EI 15.4	BE 1.98	AI 2.27	EI 2.23	BE 2.23	A 2.38	EI 2.11
	AD 16.6	AI 20.8	FG 14.9	AD 1.90	A 2.20	FG 2.19	AD 2.12	AI 2.28	FG 2.06
	AC 16.3	CE 17.7	FH 12.9	AE 1.85	CE 1.80	E 2.02	AC 2.08	CE 1.99	FH 1.80
	AE 15.7	CG 15.7	E 13.0	B 1.83	CG 1.46	FH 1.91	AE 1.99	CG 1.79	E 1.78
	B 14.7	EG 15.1	GH 12.9	AC 1.78	EG 1.39	GH 1.66	B 1.86	EG 1.70	GH 1.76
	A 13.9	CI 13.4	FI 11.8	A 1.53	EI 1.29	FI 1.55	A 1.77	CI 1.62	FI 1.74
	CD 9.8	EI 12.4	GI 11.3	CD .92	CI 1.08	GI 1.41	CD 1.32	EI 1.49	GI 1.59
	DE 8.9	E 10.8	F 8.8	DE .89	E 1.08	F 1.07	DE 1.24	E 1.30	F 1.26
	CE 8.4	C 10.6	HI 8.1	CE .79	C .95	HI 1.06	CE 1.16	C 1.28	HI 1.14
	C 4.6	GI 8.9	G 7.6	D .35	GI .85	G 1.06	C .64	GI 1.11	G 1.13
	D 3.3	G 3.7	H 3.9	E .26	G .42	H .60	D .47	G .42	H .56
	E 1.5	I -2.1	I 1.4	C .22	I -.14	I .12	E .17	I -.34	I .08
Var.	36.44	57.54	21.65	.52	.68	.50	.57	.71	.42
S.D.	6.66	7.59	4.65	.72	.82	.71	.75	.84	.65

Zero Points

H	N	L	H	N	L	H	N	L
ab 5.87	ae 7.30	ef 3.82	ab .959	ac .592	ef .717	ab .713	ae .940	ef .575
ac 2.13	ac 5.66	eg 3.02	be .102	ae .541	eg .429	ac .331	ac .736	eg .498
bc 1.42	ce 3.64	fg 1.48	bd .046	ce .233	eh .182	bc .259	ce .590	fg .330
ad .61	ag 2.56	eh 1.13	ad-.030	eg .105	gh .008	ad .123	ag .271	eh .177
bd -.12	eg -.64	fh -.26	bc-.030	ag .026	fg-.060	bd-.021	eg .022	fh .023
ae -.27	ai-1.25	ei -.97	ac-.034	cg-.092	ei-.091	ae-.049	cg-.080	gh-.069
be-1.39	cg-1.42	gh-1.36	ae-.068	ai-.210	gi-.229	be-.201	ai-.242	ei-.252
cd-1.86	ei-3.65	fi-1.67	de-.286	ci-.266	fh-.243	cd-.209	ei-.536	gi-.378
ce-2.33	ci-4.93	gi-2.30	ce-.308	ei-.353	hi-.342	ce-.349	ci-.674	fi-.396
de-4.06	gi-7.26	hi-2.89	cd-.348	gi-.581	fi-.369	de-.599	gi-1.030	hi-.500

Additivity Index

	SDc	SDo	Index	SDc	SDo	Index	SDc	SDo	Index
H	4.44	2.62	1.69	.552	.352	1.57	.532	.356	1.49
N	5.35	4.46	1.20	.653	.360	1.81	.557	.610	.91
L	2.90	2.16	1.34	.491	.333	1.47	.388	.366	1.06
Means	4.23	3.08	1.37	.565	.348	1.62	.492	.444	1.11

Combined Average Additivity Index 1.37

Context Effect Index

Var.	.64	.91.	.81
S.D.	.79	.96	.90

GIFTS

Ratio Scale Values and 0-point Deviations from *mean* 0-point

Scale Values

Means			Paired Comparisons			Multiple Standards		
H	N	L	H	N	L	H	N	L
AB 16.1	AE 18.4	EI 17.4	AB 1.74	AE 1.96	EI 2.04	AB 1.81	AE 2.15	EI 2.12
AD 13.4	AC 17.7	EG 16.7	BD 1.61	AC 1.93	EH 1.92	AD 1.53	AC 2.06	EG 2.04
BD 12.9	AI 14.7	EH 15.5	AD 1.53	AI 1.90	EG 1.88	BD 1.49	AI 1.94	EH 1.86
AE 12.0	AG 14.2	EF 14.8	AE 1.41	AG 1.74	EF 1.82	AE 1.41	AG 1.74	EF 1.75
BE 11.9	A 13.6	E 14.4	AC 1.40	CE 1.63	E 1.75	BE 1.33	A 1.68	E 1.71
BC 11.5	CE 13.6	GI 9.1	BC 1.34	A 1.62	GI 1.20	BC 1.31	CE 1.66	GI 1.02
AC 11.4	EG 9.6	GH 8.5	BE 1.21	EG 1.18	GH .92	AC 1.26	EG 1.16	GH .96
CD 9.6	CG 9.5	HI 7.6	DE 1.19	EI 1.15	HI .90	DE 1.05	EI 1.14	HI .85
DE 8.5	EI 9.4	FI 7.6	CD 1.16	CI 1.09	FG .75	CD 1.04	CI 1.12	FI .83
B 8.5	CI 9.3	FG 6.4	A 1.05	CG 1.06	FI .72	B .96	CG 1.11	G .67
A 8.3	E 8.9	FH 6.0	B .98	E .92	I .59	A .94	E 1.04	FG .66
CE 6.8	C 7.4	G 5.4	CE .83	C .87	G .58	D .74	C .87	FH .58
D 6.5	GI 3.9	I 4.6	D .67	GI .53	FH .57	CE .72	GI .37	I .47
C 3.1	I 0.7	H 3.0	C .39	I .13	H .27	C .42	I .08	H .35
E 2.3	G -.5	F .1	E .26	G -.01	F -.01	E .19	G -.06	F -.04
Var.								
13.43	22.91	26.81	.17	.37	.41	.18	.44	.42
S.D.								
3.66	4.79	5.17	.41	.61	.64	.42	.66	.65

Zero Points

H	N	L	H	N	L	H	N	L
bd 2.08	ae 4.09	eg 3.07	ab .287	ae .587	eg .449	bd .197	ae .573	eg .342
ad 1.47	ac 3.31	eh 1.81	ad .190	ac .562	ei .295	ad .139	ac .483	eh .201
ab .74	ce 2.69	ei 1.55	ac .042	ce .172	eh .103	cd .114	ce .258	gi .119
bc .13	ei .12	gi .84	be .041	ci-.083	gi-.627	ac .094	ei-.012	ei .061
ac .03	ai -.45	hi .00	bd .040	ei-.093	hi-.039	ab .085	ag-.117	gh .059
cd .03	ag-1.03	gh -.18	bc .036	ag-.128	gh-.066	bc .068	ci-.172	hi-.021
de -.68	eg-1.24	ef -.31	ae-.100	ai-.143	ef-.085	ce-.101	eg-.182	fg-.025
be-1.03	ci-1.18	fg -.96	cd-.104	cg-.203	fi-.142	de-.123	ai-.182	ef-.077
ae-1.37	cg-2.55	fi-2.91	ce-.174	eg-.258	fg-.180	be-.184	cg-.302	fh-.269
ce-1.38	gi-3.74	fh-2.91	de-.258	gi-.408	fh-.309	ae-.285	gi-.352	fi-.391

Additivity Index

	SDc	SDo	Index	SDc	SDo	Index	SDc	SDo	Index
H	2.38	1.11	2.14	.245	.156	1.57	.288	.153	1.88
N	4.24	2.44	1.74	.456	.317	1.44	.529	.309	1.71
L	4.33	1.84	2.35	.549	.215	2.55	.568	.203	2.80
Means									
	3.65	1.80	2.03	.417	.229	1.82	.462	.222	2.08

Combined Average Additivity Index 1.98

Context Effect Index

	Means	Paired Comparisons	Multiple Standards
Var.	1.16	1.15	1.11
S.D.	1.09	1.10	1.08

GIFTS
(weighted)

Ratio Scale Values and 0-point Deviations from mean 0-point

Scale Values

Paired Comparisons

H		N		L	
AB	3.13	AE	3.50	EI	3.68
BD	2.86	AC	3.43	EH	3.44
AD	2.71	AI	3.33	EG	3.38
AE	2.49	AG	3.04	EF	3.26
AC	2.48	CE	2.83	E	2.80
BC	2.34	A	2.59	GI	2.06
BE	2.25	EG	1.96	GH	1.54
DE	2.04	EI	1.94	HI	1.51
CD	2.00	CI	1.82	FG	1.27
A	1.76	CG	1.76	FI	1.17
B	1.66	E	1.55	I	1.02
CE	1.34	C	1.47	G	1.01
D	1.19	GI	.77	FH	.89
C	.71	I	.40	H	.57
E	.60	G	.08	F	.15
Var.	.54		1.10		1.27
S.D.	.73		1.04		1.12

Zero Points

H	N	L
ab .29	ae .64	eg .43
ad .24	ac .63	ei .14
bc .03	ce .19	hi .08
be .01	ci .06	gh .04
bd-.01	ei .01	fi .00
ac-.01	cg-.21	gi-.03
ce-.03	gi-.29	eh-.07
cd-.10	ai-.33	fg-.11
ae-.13	eg-.33	fh-.17
dc-.25	ag-.37	ef-.31

Additivity Index

	SDc	SDo	Index
H	.48	.15	3.14
N	.86	.38	2.26
L	1.04	.19	5.47
Means	.79	.24	3.29

Context Effect Index

Var. .84

S.D. .94

GIFTS

(Case III)

Ratio Scale Values and O-point Deviations from mean O-point

Scale Values

Paired Comparisons

H		N		L	
AB	2.50	AE	2.76	EI	2.80
BD	2.41	AC	2.68	EH	2.65
AD	2.17	CE	2.67	EG	2.60
AE	2.02	AI	2.58	EF	2.45
AC	2.01	AG	2.43	E	2.40
BC	1.95	A	2.20	GI	1.68
BE	1.88	EG	1.64	HI	1.23
DE	1.71	EI	1.63	GH	1.22
CD	1.63	CI	1.59	FI	1.02
A	1.52	CG	1.54	FG	1.00
B	1.41	E	1.23	I	.82
CE	1.19	C	1.15	FH	.80
D	.97	GI	.85	G	.75
C	.52	I	.51	H	.32
E	.44	G	.01	F	.08
Var.	.36		.68		.77
S.D.	.60		.82		.88

Zero Points

H		N		L	
ab	.43	ac	.67	eg	.55
ad	.32	ae	.67	ei	.41
ac	.03	ai	.12	eh	.06
bc	-.01	ei	.11	ef	.02
bd	-.02	ci	.07	hi	-.09
be	-.03	ag	-.23	gi	-.11
ae	-.05	ce	-.29	fi	-.13
cd	-.14	gi	-.33	gh	-.15
ce	-.22	cg	-.38	fg	-.18
de	-.29	eg	-.43	fh	-.41

Additivity Index

	SDe	SDo	Index
H	.362	.211	1.72
N	.630	.387	1.63
L	.753	.271	2.78
Means	.582	.290	2.01

Context Effect Index

Var. 1.08
S.D. 1.05

GIFTS

Ratio Scale Values and O-point Deviations from mean O-point

Scale Values

Raw Means

GOOD JUDGES

	H		N		L
AB	22.3	AE	23.4	EI	20.3
AD	19.2	AC	22.5	EG	20.0
AE	17.8	A	18.9	EH	18.9
BD	16.8	AG	18.3	E	17.8
AC	16.7	AI	18.1	EF	17.7
BC	15.9	CE	15.8	GI	9.4
BE	15.4	CG	10.9	GH	9.2
A	13.6	EG	10.3	HI	7.6
CD	12.5	E	10.1	FG	6.9
DE	12.3	EI	9.9	FI	6.3
B	10.7	CI	9.7	G	6.2
CE	8.6	C	8.6	FH	5.6
D	7.8	GI	3.2	I	3.8
E	4.0	G	-.7	H	3.6
C	3.4	I	-1.3	F	.9
Var.	27.72		54.89		42.02
S.D.	5.26		7.41		6.63

POOR JUDGES

	H		N		L
AB	9.6	AE	13.2	EI	14.1
BD	9.4	AC	12.9	EG	13.1
BE	8.2	CE	12.4	EH	11.8
BC	8.2	AI	11.1	EF	11.7
AD	7.3	AG	10.1	E	10.8
CD	6.5	EG	9.0	FI	9.4
DE	6.5	EI	9.0	GI	8.7
AE	6.4	CI	8.9	GH	7.7
B	6.3	CG	8.8	HI	7.4
AC	6.0	A	8.4	FH	6.7
D	5.0	E	7.6	FG	6.4
A	2.9	C	6.3	I	5.0
C	2.8	GI	4.8	G	4.4
CE	2.6	I	2.7	H	1.9
E	.6	G	-.2	F	1.5
	6.53		13.00		13.85
	2.55		3.60		3.72

Zero Points

	H		N		L
ad	2.1	ae	5.4	eg	4.0
ab	2.0	ac	4.9	eh	2.4
bd	1.7	ce	2.9	ei	1.3
ac	.4	ag	-.2	gi	.6
ae	-.2	ai	-.5	gh	.5
de	-.5	cg	-.9	hi	-.3
be	-.7	ei	-1.1	ef	-.8
ce	-1.2	ci	-2.4	fg	-1.6
cd	-1.3	cg	-3.0	fh	-2.9
bc	-1.8	gi	-5.2	fi	-3.4

	H		N		L
bd	1.9	ae	2.8	eh	3.4
cd	1.3	ce	2.5	eg	2.1
bc	.9	ac	1.8	ei	1.7
ce	.8	ei	1.3	gi	.7
ad	.6	ci	.1	ef	.6
ac	-.3	ai	.0	hi	-.5
ab	-.4	eg	-1.6	fg	-.5
de	-.9	ag	-1.9	gh	-1.4
be	-1.3	gi	-2.3	fi	-2.9
ae	-2.9	cg	-2.7	fh	-3.3

Additivity Index

	SDc	SDo	Index	SDc	SDo	Index
H	3.67	1.36	2.70	1.91	1.38	1.38
N	6.12	3.25	1.88	2.39	1.94	1.23
L	5.88	2.17	2.71	2.64	2.03	1.30
Means	5.22	2.26	2.31	2.31	1.78	1.30

Context Effect Index

Var.	.77	1.63
S.D.	.88	1.29

GIFTS

Ratio Scale Values and 0-point Deviations from mean 0-point

Scale Values

Multiple Standards

GOOD JUDGES			POOR JUDGES		
H	N	L	H	N	L
AC 2.78	AE 4.18	FI 2.79	AC .81	CE 1.44	EI 1.83
AB 2.31	EI 3.68	FG 2.73	BD .77	AE 1.41	EG 1.71
BE 2.08	AC 3.54	EF 2.56	AE .63	CI 1.31	EH 1.53
BD 1.94	CI 3.52	GI 2.38	AD .61	AI 1.25	EF 1.40
BC 1.87	CE 3.18	EG 2.33	AB .54	AC 1.14	E 1.39
AD 1.75	AI 3.17	G .86	BE .44	E .97	FI 1.15
AE 1.67	CG 2.54	I .81	C .43	I .95	GI 1.03
CE 1.61	E 2.53	HI .56	A .41	AG .92	GH .99
A 1.38	GI 2.50	E .52	BC .35	EI .83	HI .91
C 1.35	I 2.46	EI .51	CD .30	CG .82	FH .79
CD 1.24	AG 2.42	EH .42	E .27	GI .80	FG .73
B .78	EG 2.23	FH .18	B .25	EG .63	G .58
E .65	A 1.26	F .01	DE .07	A .45	I .56
D .19	G .78	H -.03	CE .02	G .24	H .25
DE .14	C .71	GH -.80	D -.23	C .04	F .24
Var.					
.53	.96	1.28	.07	.16	.23
S.D.					
.73	.98	1.13	.26	.40	.48

Zero Points

H	N	L	H	N	L
bd .967	ac 1.575	ef 1.033	bd .75	ac .651	fi .354
be .649	cg 1.047	fi .970	ad .43	cg .549	fh .308
ad .184	ae .389	fg .860	cd .10	ce .436	gh .155
ab .144	ag .375	fh .212	de .03	ci .328	hi .101
ac .049	ci .354	eg -.048	ac-.02	ag .229	fg-.086
bc-.260	ce -.058	eh -.060	ae-.05	ae-.010	eh-.109
cd-.296	ai -.557	hi -.226	ab-.11	ai-.151	ei-.117
ae-.356	gi -.739	gi -.295	be-.11	gi-.380	gi-.117
ce-.384	eg-1.078	ei -.820	bc-.33	eg-.573	ef-.224
de-.698	ei-1.304	gh-1.626	ce-.68	ei-1.084	eg-.265

Additivity Index

	SDc	SDo	Index	SDc	SDo	Index
H	.664	.479	1.39	.257	.374	.69
N	.616	.883	.70	.282	.519	.54
L	1.25	.793	1.58	.374	.205	1.82
Means						
	.84	.72	1.17	.30	.37	.81

Context Effect Index

Var.	.92	1.40
S.D.	.97	1.21

Paired Comparisons Matrix P for High Context Behavior

							Items								Items
AB	AC	AE	AD	BD	BC	BE	A	CD	B	DE	CE	D	C	E	
--								.810		.820	.805	.930	.910	.930	AB
	--			.645		.610			.745	.705		.840		.890	AC
		--		.595	.520			.695	.700			.870	.820		AE
			--		.550	.550			.690		.725		.855	.890	AD
				--			.590				.700		.815	.915	BD
					--		.540			.655		.805		.860	BC
						--	.550	.595				.800	.770		BE
							--	.525	.590	.565	.640	.775	.775	.860	A
								--	.520					.870	CD
									--	.515	.595	.760	.745	.845	B
										--			.745		DE
											--	.665			CE
												--	.540	.565	D
													--	.645	C
														--	E

Multiple Standards Cumulative Matrix P for
High Context Behavior Upper Category Boundary

	40	45	50	55	60	65	70
AB	.000	.000	.025	.135	.310	.720	.880
AC	.000	.000	.015	.145	.390	.795	.910
AD	.000	.000	.015	.155	.390	.840	.925
AE	.000	.020	.050	.190	.395	.870	.915
BC	.005	.030	.070	.225	.460	.815	.905
BE	.000	.020	.075	.245	.515	.865	.925
BD	.000	.030	.095	.345	.575	.880	.935
CE	.000	.045	.140	.450	.660	.915	.955
CD	.000	.025	.120	.425	.700	.930	.955
A	.005	.090	.165	.460	.660	.950	.980
DE	.000	.060	.160	.505	.755	.935	.955
B	.000	.075	.175	.540	.725	.950	.975
C	.005	.135	.315	.680	.845	.995	1.000
D	.000	.130	.365	.755	.895	.995	1.000
E	.005	.360	.565	.870	.950	.995	1.000

Paired Comparisons Matrix P for Neutral Context Behavior

Items

AC	AE	AG	A	AI	CE	CG	CI	C	EG	EI	E	GI	G	I	Items
--									.910	.930	.955	.970	.980	.980	AC
	--					.865	.855	.860				.970	.980	.995	AE
		--			.555		.760	.810		.910	.920			.995	AG
			--		.590	.730		.795	.815	.880	.885	.955	.975	.985	A
				--	.570	.695		.795	.820		.885		.970		AI
					--							.950	.960	.980	CE
						--				.760	.755			.980	CG
							--		.710		.725		.905		CI
								--	.605	.675	.670	.865	.890	.910	C
									--					.945	EG
										--			.845		EI
											--	.745	.830	.915	E
												--			GI
													--	.705	G
														--	I

Multiple Standards Cumulative Matrix P for Neutral

Context Behavior Upper Category Boundary

	40	45	50	55	60	65	70
AC	.000	.015	.015	.065	.225	.580	.790
AE	.000	.015	.025	.080	.265	.725	.870
AG	.000	.015	.040	.210	.435	.880	.950
CE	.000	.020	.045	.220	.465	.885	.935
AI	.000	.030	.055	.295	.520	.930	.985
A	.000	.045	.075	.290	.530	.910	.955
CG	.000	.020	.075	.385	.685	.935	.975
CI	.005	.020	.090	.390	.635	.945	.980
C	.015	.090	.145	.435	.675	.945	.980
EG	.000	.120	.270	.680	.880	.995	.995
EI	.000	.170	.365	.750	.925	.990	.995
E	.025	.225	.430	.750	.885	.990	.995
GI	.030	.335	.585	.910	.975	1.000	1.000
G	.030	.485	.740	.995	1.000	1.000	1.000
I	.065	.700	.860	1.000	1.000	1.000	1.000

Paired Comparison Matrix P for Low Context Behavior

Items

EF	EH	EG	EI	E	FG	GH	FH	FI	HI	F	GI	H	G	I	Items
--						.810			.815		.885	.860	.935	.945	EF
	--				.690			.770		.830	.880		.910	.970	EH
		--					.655	.765	.765	.795		.810		.945	EG
			--		.635	.640	.620			.735		.790	.860		EI
				--	.665	.630	.620	.670	.705	.720	.820	.765	.840	.940	E
					--				.560			.695		.870	FG
						---		.545		.645				.895	GH
							--				.615		.675	.805	FH
								--				.610	.630		FI
									--	.575			.590		HI
										--	.515	.490	.575	.710	F
											--	.545			GI
												--	.475	.640	H
													--	.705	G
														--	I

Multiple Standards Cumulative Matrix P for Low Context Behavior Upper Category Boundary

	40	45	50	55	60	65
EF	.000	.020	.120	.375	.625	.870
EG	.010	.045	.095	.355	.625	.890
EH	.010	.085	.165	.440	.670	.890
EI	.020	.125	.220	.520	.780	.965
E	.010	.095	.195	.555	.825	.965
FG	.010	.105	.280	.670	.835	.975
FI	.000	.080	.310	.725	.885	.975
FH	.020	.115	.360	.685	.855	.975
GH	.010	.155	.395	.725	.880	.975
F	.020	.240	.470	.785	.885	1.000
GI	.010	.255	.455	.825	.950	1.000
HI	.025	.250	.495	.840	.950	1.000
G	.035	.310	.495	.860	.965	1.000
H	.040	.465	.635	.865	.945	1.000
I	.035	.595	.810	1.000	1.000	1.000

Paired Comparison Matrix P for *High* Context Violations

Items

Items	AB	BD	BC	BE	AD	AE	B	AC	A	CD	DE	CE	D	E	C
AB	--									.935	.950	.960	.970	.985	.985
BD		--				.620		.665	.745			.915		.960	.970
BC			--		.590	.630			.745		.855		.970	.950	
BE				--	.605			.590	.690	.855			.935		.975
AD					--		.480					.870		.960	.970
AE						--	.410			.890			.950		.935
B							--	.545	.655	.800	.810	.870	.880	.935	.955
AC								--			.835		.955	.905	
A									--	.765	.720	.895	.890	.915	.900
CD										--				.810	
DE											--				.700
CE												--	.765		
D													--	.545	.585
E														--	.460
C															--

Multiple Standards Cumulative Matrix P for High Context Violations Upper Category Boundary

	40	45	50	55	60	65	70	75
AB	.000	.000	.005	.020	.050	.360	.610	.710
BD	.000	.000	.000	.030	.175	.585	.765	.865
BC	.000	.005	.000	.090	.200	.565	.765	.860
BE	.000	.000	.000	.065	.225	.575	.765	.880
AD	.000	.000	.025	.100	.210	.640	.835	.870
AC	.005	.015	.015	.009	.220	.660	.850	.890
AE	.000	.000	.000	.145	.265	.710	.835	.880
B	.000	.015	.020	.165	.280	.730	.880	.925
A	.015	.020	.050	.225	.400	.770	.860	.905
CD	.000	.020	.065	.310	.630	.915	.960	.995
DE	.000	.030	.090	.365	.655	.915	.960	1.000
CE	.000	.030	.130	.415	.745	.905	.955	.975
C	.015	.130	.295	.675	.820	.950	.975	1.000
D	.025	.180	.335	.700	.885	.985	1.000	1.000
E	.040	.275	.460	.825	.910	.970	.995	1.000

Paired Comparisons Matrix P for Neutral Context Violations

Items

AE	AG	AC	AI	A	CE	CG	EG	EI	E	CI	C	GI	G	I	Items
--						.885				.945	.950	.985	.995	1.000	AE
	--				.805			.945	.960	.920	.925			1.000	AG
		--					.880	.910	.935			.960	.985	.980	AC
			--		.740	.785	.780		.880		.920		.960		AI
				--	.710	.785	.780	.810	.875	.875	.900	.900	.955	.995	A
					--							.800	.925	.985	CE
						--		.615	.630					.970	CG
							--			.605	.650			1.000	EG
								--			.615		.885		EI
									--	.505	.485	.630	.810	.980	E
										--			.705		CI
											--	.515	.660	.810	C
												--			GI
													--	.825	G
														--	I

Multiple Standards Cumulative Matrix P for Neutral Context Violations Upper Category Boundary

	40	45	50	55	60	65	70	75
AC	.000	.000	.000	.030	.100	.360	.575	.750
AE	.000	.000	.000	.040	.125	.495	.670	.775
AG	.000	.000	.000	.045	.160	.570	.740	.880
A	.005	.015	.020	.085	.230	.660	.770	.850
AI	.000	.000	.000	.080	.300	.685	.830	.860
CE	.000	.020	.075	.195	.400	.700	.855	.910
CG	.000	.025	.085	.280	.550	.840	.885	.925
EG	.000	.015	.075	.370	.605	.860	.900	.930
CI	.015	.105	.155	.375	.585	.860	.905	.945
EI	.020	.060	.155	.480	.670	.905	.945	.960
E	.005	.145	.230	.555	.725	.920	.955	.970
C	.025	.170	.220	.525	.695	.945	.955	.970
GI	.005	.125	.275	.725	.885	.945	.955	.970
G	.060	.385	.555	.895	.955	.975	.985	.995
I	.220	.715	.870	.955	.780	.995	.995	.995

Paired Comparison Matrix P for Low Context Violations

Items

EG	EH	EF	EI	FG	E	FH	GH	FI	GI	F	G	HI	H	I	Items
—						.805		.910		.970		.900	.965	.980	EG
	--							.890	.790	.995	.905			.980	EH
		--		.620			.805		.870		.890	.910	.960	.970	EF
			--	.535		.680	.695			.980	.880		.940		EI
				--	.650							.830	.935	.975	FG
					--	.495	.590	.625	.695	.965	.835	.825	.945	.945	E
						--			.640		.760			.970	FH
							--	.525		.735				.945	GH
								--			.655		.880		FI
									--	.535			.760		GI
										--	.560	.645	.770	.920	F
											--	.525	.620	.830	G
												--			HI
													--	.710	H
														--	I

Multiple Standards Cumulative Matrix P for Low Context Violations Upper Category Boundary

	40	45	50	55	60	65	70
EF	.000	.020	.040	.230	.345	.750	.865
EG	.000	.015	.050	.195	.435	.765	.860
EH	.000	.015	.060	.295	.505	.835	.935
EI	.000	.015	.065	.280	.525	.830	.960
FG	.000	.030	.120	.325	.540	.860	.915
FH	.005	.055	.175	.460	.660	.885	.950
E	.005	.060	.140	.430	.600	.925	.965
GH	.000	.040	.170	.465	.705	.925	.950
FI	.000	.030	.190	.510	.725	.925	.985
GI	.000	.075	.240	.525	.770	.920	.960
F	.005	.175	.320	.625	.830	.965	.995
HI	.005	.150	.440	.725	.865	.970	.980
G	.025	.235	.395	.710	.835	.965	.970
H	.055	.425	.640	.850	.945	.980	.995
I	.110	.600	.790	.945	.975	.985	.985

Paired Comparisons Matrix P for High Context Gifts

AB	BD	AD	AE	AC	BC	BE	DE	CD	A	B	CE	D	C	E	Items
--							.735	.750			.860	.845	.905	.925	AB
	--		.530	.540					.660		.810		.920	.925	BD
		--			.590	.605				.740	.765		.880	.875	AD
			--		.575			.600		.695		.765	.810		AE
				--		.565	.630			.660		.790		.825	AC
					--		.560		.635			.700		.895	BC
						--		.570	.620			.710	.835		BE
							--		.525	.590			.835		DE
								--	.535	.600				.810	CD
									--	.560	.565	.645	.760	.760	A
										--	.580	.670	.745	.780	B
											--	.590			CE
												--	.625	.640	D
													--	.535	C
														--	E

Multiple Standards Cumulative Matrix P for High Context Gifts Upper Category Boundary

	40	45	50	55	60	65	70
AB	.010	.060	.070	.235	.395	.695	.845
AD	.020	.070	.125	.335	.480	.805	.920
BD	.030	.075	.130	.400	.565	.790	.890
AE	.025	.075	.135	.380	.575	.855	.930
BE	.010	.080	.170	.430	.625	.835	.935
BC	.010	.075	.160	.430	.655	.866	.935
AC	.030	.095	.165	.395	.600	.895	.955
DE	.020	.095	.195	.525	.735	.930	.980
CD	.020	.110	.260	.530	.670	.920	.965
B	.020	.165	.245	.590	.720	.940	.955
A	.035	.225	.290	.525	.665	.945	.960
D	.030	.205	.365	.700	.810	.940	.980
CE	.040	.165	.335	.680	.830	.970	.990
C	.080	.380	.555	.765	.855	.960	.975
E	.035	.435	.525	.870	.935	1.000	1.000

Paired Comparisons Matrix P for Neutral Context Gifts

AE	AC	AI	AG	A	CE	EG	EI	CI	CG	E	C	GI	I	G	Items
--								.880	.865		.910	.880	.940	.970	AE
	--					.865	.815			.895		.900	.950	.950	AC
		--			.610	.795			.805	.855	.850			.965	AI
			--		.600		.755	.740		.840	.795		.915		AG
				--	.565	.740	.740	.715	.725	.785	.790	.815	.885	.940	A
					--							.885	.940	.990	CE
						--		.580			.685		.890		EG
							--		.555		.640			.920	EI
								--		.585				.925	CI
									--	.555			.865		CG
										--	.550	.725	.800	.870	E
											--	.720	.780	.870	C
												--			GI
													--	.525	I
														--	G

Multiple Standards Cumulative Matrix P for Neutral Context Gifts Upper Category Boundary

	40	45	50	55	60	65	70
AE	.010	.050	.100	.240	.360	.700	.850
AC	.005	.050	.105	.260	.405	.745	.885
AI	.020	.035	.110	.365	.535	.815	.890
AG	.005	.055	.140	.350	.575	.895	.955
A	.020	.105	.175	.365	.550	.895	.940
CE	.000	.050	.160	.430	.635	.920	.945
EG	.020	.125	.240	.675	.820	.970	1.000
EI	.020	.125	.285	.690	.810	.970	.995
CI	.025	.135	.270	.685	.850	.965	.990
CG	.035	.140	.295	.660	.850	.955	.980
E	.030	.205	.310	.700	.815	.970	.980
C	.075	.240	.385	.735	.860	.970	.990
GI	.110	.420	.650	.885	.945	.980	.995
I	.180	.630	.760	.930	.945	.990	.990
G	.220	.660	.790	.955	.990	1.000	1.000

Paired Comparison Matrix P for Low Context:Gifts

Items

HI	H	GH	A	GI	EH	EF	FI	FG	EG	FH	EI	F	E	G	Items
--						.135		.555	.130			.880	.180	.610	HI
	--		.420	.180		.075	.270	.290	.040		.040	.655	.060	.375	H
		--	.690			.175	.625				.085	.840	.150		GH
			--		.115	.125		.380	.085	.455		.805	.130	.455	A
				--	.205	.205				.760		.925	.240		GI
					--		.880	.895				.950		.955	EH
						--								.905	EF
							--		.115				.175	.495	FI
								--			.100		.150		FG
									--	.915		.940			EG
										--	.060		.140	.475	FH
											--	.945		.960	EI
												--	.065	.215	F
													--	.905	E
														--	G

Multiple Standards Cumulative Matrix P for Low .

Context, Gifts Upper Category Boundary

	40	45	50	55	60	65	70
EI	.000	.040	.080	.310	.445	.830	.920
EG	.025	.055	.100	.330	.515	.820	.925
EH	.010	.055	.115	.395	.585	.900	.955
EF	.010	.075	.140	.410	.625	.920	.975
E	.020	.105	.150	.375	.595	.940	.975
GI	.020	.215	.430	.745	.885	.960	.980
GH	.045	.235	.425	.750	.915	1.000	.990
HI	.040	.280	.510	.825	.915	.965	.990
FI	.060	.290	.525	.805	.905	.965	.975
G	.100	.375	.555	.850	.925	.990	.990
FG	.100	.355	.515	.840	.955	.990	1.000
FH	.085	.355	.615	.870	.960	1.000	1.000
I	.095	.525	.705	.895	.925	.975	.975
H	.135	.545	.720	.925	1.000	1.000	1.000
F	.260	.700	.810	1.000	1.000	1.000	1.000

Multiple Standards Cumulative Matrix P for High Context
Gifts - Good Judges Upper Category Boundary

	40	45	50	55	60	65	70
A	.000	.024	.153	.459	.718	.929	.988
B	.024	.129	.353	.741	.894	.988	1.000
C	.012	.024	.188	.506	.682	.929	.965
D	.024	.435	.553	.871	.953	1.000	1.000
E	.012	.129	.388	.706	.894	.965	.977
AB	.000	.000	.000	.129	.318	.718	.877
AC	.000	.000	.000	.035	.188	.553	.788
AD	.012	.035	.141	.329	.600	.788	.906
AE	.000	.012	.094	.341	.659	.847	.941
BC	.000	.012	.059	.235	.541	.835	.929
BD	.012	.024	.082	.318	.482	.753	.871
BE	.000	.000	.035	.200	.424	.788	.894
CD	.012	.106	.188	.553	.718	.953	.977
CE	.000	.059	.106	.341	.577	.929	.941
DE	.059	.424	.624	.918	.929	1.000	1.000

Multiple Standards Cumulative Matrix P for Neutral Context

Gifts - Good Judges Upper Category Boundary

	40	45	50	55	60	65	70
A	.136	.531	.827	.963	.988	.988	1.000
C	.296	.778	.914	1.000	1.000	1.000	1.000
E	.012	.111	.284	.765	.877	.975	1.000
G	.272	.753	.901	1.000	1.000	1.000	1.000
I	.000	.111	.321	.790	.889	.963	1.000
AC	.000	.000	.037	.198	.420	.877	.951
AE	.000	.000	.025	.074	.210	.654	.815
AG	.012	.161	.296	.728	.926	.988	.988
AI	.000	.037	.123	.395	.630	.901	.926
CE	.000	.062	.148	.370	.531	.877	.914
CG	.037	.136	.284	.691	.889	.963	.963
CI	.000	.000	.037	.210	.444	.877	.951
EG	.037	.198	.370	.790	.951	.988	1.000
EI	.000	.000	.037	.161	.407	.840	.901
GI	.012	.148	.259	.741	.889	1.000	1.000

Multiple Standards Cumulative Matrix P for Low Context

Gifts - Good Judges Upper Category Boundary

	40	45	50	55	60	65	70
E	.013	.253	.573	.920	.987	1.000	1.000
F	.120	.560	.733	.960	1.000	1.000	1.000
G	.013	.240	.467	.760	.933	.987	1.000
H	.120	.533	.800	.960	.987	.987	.987
I	.027	.200	.467	.787	.960	.973	.987
EF	.000	.013	.027	.267	.453	.827	.933
EG	.000	.000	.040	.280	.520	.893	.973
EH	.080	.333	.667	.987	.987	.987	.987
EI	.067	.373	.560	.867	.987	1.000	1.000
FG	.000	.000	.027	.240	.387	.747	.880
FH	.093	.400	.720	.947	1.000	1.000	1.000
FI	.000	.000	.027	.173	.320	.760	.893
GH	.307	.827	.933	1.000	1.000	1.000	1.000
GI	.000	.013	.053	.267	.480	.907	.947
HI	.053	.360	.573	.867	.960	1.000	1.000

Multiple Standards Cumulative Matrix P for High Context

Gifts - Poor Judges Upper Category Boundary

	40	45	50	55	60	65	70
A	.036	.167	.238	.595	.762	.929	.976
B	.036	.286	.382	.667	.726	.893	.964
C	.024	.202	.333	.560	.667	.905	.964
D	.048	.441	.488	.869	.917	1.000	1.000
E	.071	.202	.286	.655	.762	.976	1.000
AB	.036	.143	.250	.548	.655	.905	.964
AC	.024	.119	.143	.441	.607	.845	.917
AD	.012	.132	.202	.536	.643	.881	.964
AE	.024	.143	.226	.524	.643	.881	.941
BC	.060	.179	.274	.560	.667	.964	.976
BD	.048	.132	.179	.488	.643	.821	.905
BE	.048	.155	.238	.571	.738	.929	.976
CD	.024	.226	.310	.619	.714	.929	.929
CE	.071	.393	.476	.714	.762	.964	.976
DE	.107	.333	.476	.714	.774	.917	.952

Multiple Standards Cumulative Matrix P for Neutral Context

Gifts - Poor Judges Upper Category Boundary

	40	45	50	55	60	65	70
A	.086	.309	.469	.803	.901	.975	.988
C	.148	.543	.667	.904	.975	1.000	1.000
E	.025	.136	.198	.580	.765	.963	1.000
G	.086	.506	.617	.864	.889	.975	.975
I	.037	.136	.247	.593	.728	.975	.988
AC	.012	.111	.247	.506	.728	.914	.963
AE	.012	.086	.173	.432	.593	.827	.938
AG	.037	.111	.247	.642	.778	.938	.988
AI	.000	.062	.198	.469	.642	.938	.963
CE	.025	.099	.198	.432	.556	.803	.914
CG	.037	.148	.309	.630	.815	.951	.988
CI	.037	.074	.185	.519	.630	.877	.951
EG	.111	.284	.407	.691	.765	.951	.963
EI	.037	.210	.321	.580	.704	.963	.975
GI	.049	.259	.358	.654	.741	.938	.963

Multiple Standards Cumulative Matrix P for Low Context Gifts - Poor Judges Upper Category Boundary

	40	45	50	55	60	65	70
E	.041	.192	.233	.466	.685	.959	.986
F	.192	.548	.671	.917	.959	.986	1.000
G	.137	.370	.521	.836	.890	.986	1.000
H	.137	.534	.712	.890	.959	1.000	1.000
I	.082	.521	.616	.822	.863	.959	.959
EF	.014	.137	.233	.534	.726	.945	.959
EG	.055	.096	.164	.411	.644	.890	.973
EH	.014	.096	.206	.521	.699	.959	.973
EI	.000	.082	.123	.438	.562	.890	.945
FG	.110	.301	.438	.781	.890	.986	.986
FH	.082	.274	.479	.753	.877	.973	.986
FI	.041	.233	.370	.658	.795	.932	.945
GH	.069	.206	.343	.740	.890	.986	.986
GI	.014	.233	.411	.712	.808	.945	.973
HI	.069	.301	.438	.753	.849	.932	.973

For K. R. Hammond (Ed.), Probabilistic functionalism: the psychology of Egon Brunswik. New York: Holt, Rinehart & Winston, late 1964 or early 1965.

Pattern Matching as an Essential in Distal Knowing[1]

Donald T. Campbell

To establish the setting and the problem, this paper first deals with Egon Brunswik's concept of distality in terms of behavior theory and evolutionary development. Even here, knowing in science is introduced as in continuity with those modes of knowing on the part of organisms described by biologists and psychologists. The problem is then introduced: How is distal knowing achieved? How are the proximal threads tied in a distal knot? The thesis is advanced that pattern matching is one of the recurrent attributes. This is illustrated first at the level of stereoscopic perception of objects and the coordination of motor behavior. Further illustrations come from the diagnosis of inferred entities in science. Finally, the perspective is used to summarize an emerging consensus among philosophers of science as to the relationship between theory and data.

Joining analytic philosophers since Hume, both scientific knowledge and ordinary knowledge of the common-sense objects of the external world are recognized as analytically unjustified, highly presumptuous, and fallible. This recognition seems basic to both philosophical and psychological approaches to the problem of knowledge. But if to the problem of knowledge, the philosopher answers in a specific instance or in general that "knowledge is impossible," he has answered a different question than was asked. For the question, "How is knowledge possible?" usually implies

---

[1] Supported in part by U.S. Office of Education Project C-998, Contract 3-20-001, under provisions of Title VII of the National Defense Education Act.

that knowledge, corrigible though it may be, has in some instances been achieved to a sufficient degree to make the question worth asking. It presumes that instances of valid knowing and of mistake both occur and are in some instances discriminable. How are such discriminations made? How can one identify new instances as "the same" as old ones so that "knowledge" may be applied? How do knowers choose between the interpretation of sense data as indicating a new entity versus connoting an old entity from a changed perspective? It is the problem of knowledge at this level to which the present inquiry is addressed.

By distal knowledge is meant in particular knowledge of external objects and events, knowledge of predictable processes reidentifiable as the same, hypothetical knowledge optimally invariant over points of observation and observers. Perhaps in the truest connotations of the word knowledge, all knowledge is distal, and our title has a needless redundancy. Perhaps there is no sense in which punctiform raw sense data unidentified as a recognizable recurrent, uninterpreted as diagnostic of external events, should be called knowledge.

Both psychology and philosophy are emerging from an epoch in which the *quest for punctiform certainty* seemed the optimal approach to knowledge. To both Pavlov and Watson, single retinal cell activations and single muscle activations seemed more certainly reidentifiable and specifiable than perceptions of objects or adaptive acts. The effort in epistemology to remove equivocality by founding knowledge on particulate sense data, and the spirit of logical atomism, point to the same search for certainty in particulars. These are efforts of the past, now increasingly recognized to be untenable, yet the quest for punctiform certainty is still a pervasive part of our intellectual background. A preview of the line argument as it relates to the nostalgia for certainty through incorrigible particulars may be provided by the following analogy. Imagine the task of identifying "the same" dot of ink in two newspaper prints of the same photograph. The task is impossible if the photographs are examined by exposing only one dot at a time. It becomes more possible the larger the area of each print exposed. Insofar as any certainty in the identification of a single particle is achieved, it is because a prior identification of the whole has been achieved. Rather than the identification of the whole being achieved through the firm establishment of particles, the reverse is the case, the complex being more certainly known than the elements, neither, of course, being known incorrigibly.

The operational distinction between distal and proximal S-R laws

It is today fashionable in the psychology of learning to acknowledge that molar responses and molar stimuli are involved in the empirical regularities upon which behavior theories are based. But usually this is presented as the scientist's arbitrary choice of a convenient level of analysis, rather than in recognition of the empirical fact that it is at a molar level rather than at the molecular level that S-R laws can be obtained. This recognition is fundamental to Brunswik's similar distinction between the distal and proximal levels. Examination of his "lens model" in Fig. 1 will help make this clear.

Brunswik recognized the scientific legitimacy of the search for proximal s-r laws, relationships between single retinal cell activation and specific muscle contraction, for example. However, he summarized psychology's experience with the higher vertebrates by the general induction that such laws did not exist; that when computed, these correlations (between presence-absence of stimulus and presence-absence

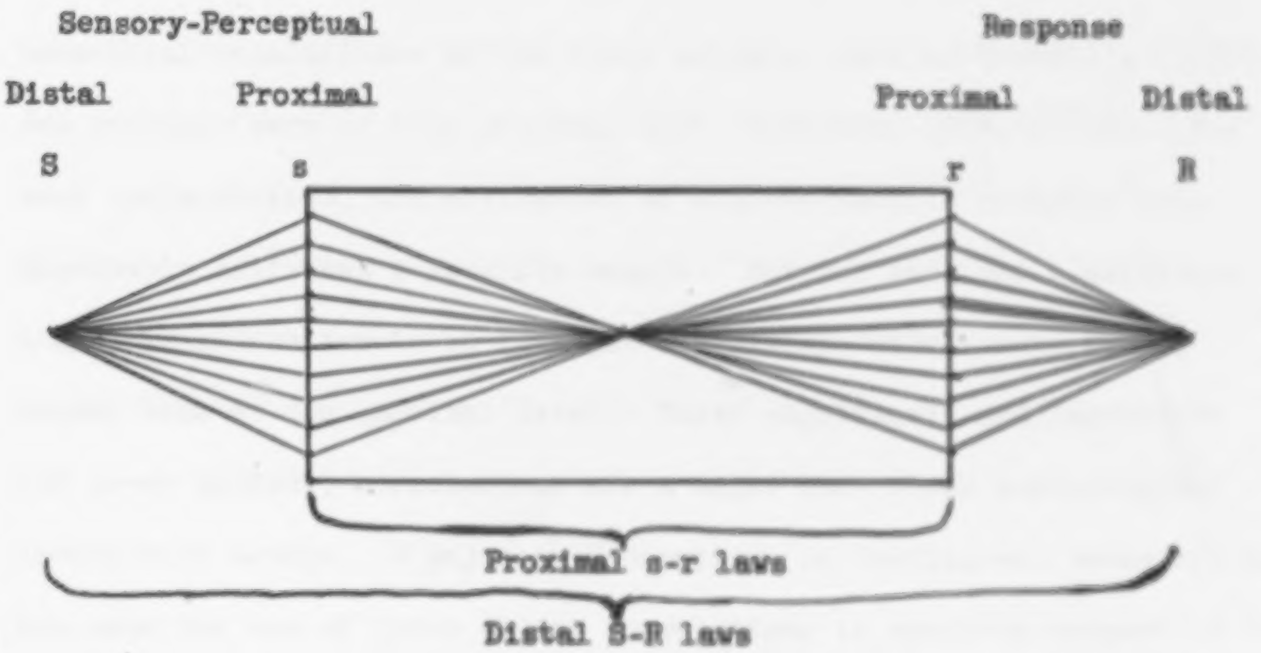

Sensory-Perceptual
Response
Distal
Proximal
Proximal
Distal
S
s
r
R
Proximal s-r laws
Distal S-R laws

Fig. 1. Brunswik's "lens model." The box represents a conceptual organism, with a sensory-perceptual surface (to the left) and a response surface (to the right). The radial lines indicate the multifarious proximal channels alternatively (vicariously) mediating the distal achievement. On the stimulus side, these could be illustrated by the multiple intersubstitutable proximal cues for distance or for object color, etc. On the response side these could be illustrated by the multiple specific muscle contractions by which the rat might depress a lever, or locomote a runway, etc.

of muscle contraction over a population of occasions) were of zero magnitude. He did not deny the possibility of discovering such laws in any programmatic or logical sense, and in fact emphasized that the behavioral regularities of the lower animals, such as Uexküll's (1934) sea urchins, were of this proximal sort (Brunswik, 1956, p. 62). For such coelenterates, the activation of a given tactile receptor cell dependably activates a specific muscle. But for man, the significant S-R correlations are to be found at the distal, object-act, level, rather than at the proximal level. These significant and impressive (if never perfect) correlations are a major fact which psychological theory must accept. A major self-deception of traditional behaviorisms has been the use of these distal correlations in specious support of theories adequate only to explain proximal ones. A major achievement of Brunswik's perspective is that of maintaining a positivistic, behaviorsitic orientation without denying the flexible, purposive adaptedness characteristic of the perceptual and motor systems of the higher organisms. That such distal foci are operationally distinguishable from the proximal has been emphasized by Smedslund (1953) and Campbell (1954) in defense of a Brunswikian point of view on the problem of "what is learned" and in summary of the research in learning showing the distal character of most learned habits.

The achievement of such a point of view meant overcoming special temptations for a person of Brunswik's positivist background. The positivist's militant attack upon vitalism and upon the metaphysical baggage of pseudoproblems made them prone to overlook or deny these less tangible facts of perception and behavior. The particular form which their quest for certainty took was in seeking a certainty of communication, i.e., in seeking unequivocally specifiable terms. Proximal stimulus and proximal responses seem much more appealing in this regard than do distal ones. It was Brunswik's achievement to make objective the facts which necessitated the renunciation of this approach to certainty, facts to which many positivists within psychology are still blind. It was his achievement to have retained both his vigorous positivism, and to have harnessed these "organismic" and "vitalistic" facts as positivistically demonstrated laws. His empirical law of distal achievements through vicarious mediation I accept as established, setting the problem for this essay.

Evolutionary perspective upon degrees of distality

In the primitive coelenterate, the nervous system connects each specific tactile sense organ with a specific muscle. The s-r reflexes are adaptive, in that the muscle response changes the relationship between the organism and environment (in some instances leading to tenacle withdrawing, in other instances, contraction and grasping, depending on location, etc.). While these are distal effects, there is but a single means of mediating them--there are no alternative channels for vicarious mediation. Thus at the proximal level, the correlation holds. Higher forms, even locomotor forms, may possibly preserve this proximal consistency. Loeb (1918) described an ingenious mechanical bug with presumed animal counterparts which flexibly tracked a light. The left eye or photocell activated the right hind leg or wheel, and the right eye the left hind leg. Thus when both eyes were equally stimulated, locomotion was straight ahead, toward the light.

If the initial orientation was to one side, the eye receiving the greater amount of light activated the opposite side the more, leading the bug to turn in the direction of the light, up to that point where the eyes were equally stimulated.

But the major stable relationships in the world available for organismic exploitation are consistencies adhering to other objects than the organism itself--consistencies adhering to food objects, shelters, predators, locomotor obstacles and throughways, etc. Due to the fluctuating illumination and the variable distance of such objects, proximal stimuli are not optimal indicators of these, nor are proximal responses optimal effectors. Somewhere in the evolutionary hierarchy the available distal relationships come to be exploited, and with this comes a renunciation of rigid one-to-one reflexes at the proximal level. Presumably this has been achieved by the evolutionary level at which image-forming eyes appear. The organism at that point tries out the strategy of hypothesizing stable external objects mediately known, or of behaving in a manner consistent with such hypothesizing.

This is awkwardly conveyed on two counts. On the one hand, the statement is atrociously anthropomorphic. On the other, we so unquestioningly assume the existence of specific external objects--objects with a stability independent of our movements--that we find it hard to accept these as merely the hypotheses of a fallible cognitive apparatus, at best mediately known. We find it hard to comprehend an organism which does not live and locomote in a world of objects. Yet this feature of animal life is clearly dependent upon distance receptors and, in any high degree, upon perceptual constancy mechanisms.

S-R consistencies, whether instinctive or habitual, are dependent for their establishment and maintenance upon consistencies in the environment, inasmuch as selective survival or selective reinforcement are essential. Initially, only environmental contingencies at the skin of the organism are involved. At higher stages, environmental contingencies at some remoteness from the skin can be diagnosed, and the diagnostic inference becomes to some degree independent of the position of the observing organism. Since it is ecologically true that these less self-centered, more "objective" contingencies are stronger statistically (provide better predictions) there is a selection-pressure in evolution leading to their exploitation, and to the increasingly complex perceptual-locomotor apparatus which makes this possible.

Students of science and epistemology have long found it appropriate to tease out the hidden premises underlying the reasoning of scientists and laymen. These premises are often for convenience stated in a language of conscious contents: the scientist presumes order (or partial order) in nature, he presumes that there are fewer causal laws than there are events, etc. These premises are unconscious until thus explicated. It is in a similar vein that I use a conscious experience terminology of "hypothesizing" in discussing the implicit presumptions about the nature of the world built into animals at various evolutionary stages (Lorenz, 1941; Campbell, 1959). These presumptions, these synthetic hypotheses (ontogenetically a priori or a posteriori), increase with the adaptive radius, the range of correspondences (Spencer, 1896) of the organism, and represent the increased "knowledge" of the laws of the world, of the "causal texture of the environment" (Tolman & Brunswik, 1935).

In these terms, the proximal-level organism in his inherited reflexes modestly assumes some degree of order, some deviation from pure randomness in his environment. At some later level, his ambitious presumptiveness goes so far as to presume the existence of external objects. At this stage, the organism's cognitive apparatus becomes a diagnoser of external entities and processes. At this stage, the puzzling fact of distality and of purposive behavior emerges in full degree. From the standpoint of an amoeba, this is a wonderous, if very presumptive, procedure.

Continuing the evolutionary paradigm, we can note that the higher the level of development the higher the degree of distality achieved, the greater the magnitude of the Brunswik constancy ratio, and the greater the degree to which external events and objects are known in a manner independent of the point of view of the observer. Thus Brunswik (1928; 1956) found the degree of constancy achieved to be higher in older children than younger ones, reaching a maximum around 14 or 15 years of age. Thus Piaget (e.g., Flavell, 1963; Piaget, 1937) finds his youngest children failing to reify external objects as having continuing existence and motion when out of sight. With increasing age come not only increasing hypostatization of external objects, but also of increasing presumption of the conservation of weight and volume, as the child more and more imputes to the world stability under transformations of his point of observation. Thus, within the historical span of human experience, the distality of man's knowledge has been increased by learning and reason over that provided by visual perception. Brunswik both emphasized the marvelous constancy mechanisms of the visual and auditory systems, and at the same time spoke of the "stupidity" of the perceptual apparatus relative to thinking and in its rigidity and susceptibility to illusion under ecologically atypical conditions (e.g., Brunswik, 1934; 1956, pp. 88-92). Piaget provides similar evidence. Man's introduction of measuring rods constructed of relatively inelastic quasi solids has freed his knowledge of objects from the ubiquitous compromise with retinal size found in unaided vision (Brunswik, 1956, p. 88 & ad passim).

The course of science has this over-all trend. The intricacies of measuring instruments and laboratory equipment shows a development analogous to that of the vertebrate eye. The addition of specific compensating devices and control features leads to meter readings to reflect more and more purely the attributes of the object of study, uncontaminated by the irrelevant specifics, vagaries, and rigidities of the measuring instrument. The concept of a compensated or of an automatically self-calibrating instrument, and the need for continually re-calibrating a more primitive one, convey the general spirit of this development. (See Wilson, 1952, for specific illustrations.)

Similarly, physical theory has provided in each generation a model of the universe more independent of man's particular position of observation. In the shift from Ptolemy to Kepler this is clear. Some see the contributions of relativity and quantum theory and the complementarity principle as reversals in this direction. This is not the position taken here, however. What these developments have done is to convince man that his knowledge of the world is not as completely distal or objective as he had thought, that it is still contaminated by a certain degree of astigmatism (Bachem, 1951), and that there are limitations upon the degree of objective knowledge which he can obtain. But note that in many specifics if not all the errors of the Newtonian system become correctible when an Einsteinian relativity is adopted, achieving a greater degree of distality, independence of point of observation, and predictive efficacy greater than that previously held. Oppenheimer (1956) has presented a parallel development in early science. He contrasts Babylonian astronomy with the later Mediterranean versions of Ptolemy and Copernicus. Babylonian astronomy was able to predict the movement of the heavenly bodies with great accuracy, including such subtleties as the eclipses of the moon, without any model of celestial mechanics. This achievement was the product of centuries of bookkeeping, and was based upon vast libraries of detail from which identical sequences and spatial contingencies could be identified as a base for prediction. The great advantage of the model-building astronomies was not one of accuracy (at least initially) but rather one of library space. The same repertoire of predictions could be encompassed with one one-thousandth of the written records. Similarly, while it is conceivable that instinct or learning provide an animal with a fixed response tendency for each possible proximal pattern of stimulation, the central nervous system storage requirements involved for an animal like ourselves with proximal stimulus receptors numbering in the millions, are such as to argue the economy of the more presumptive strategy of creating an approximate model of the environment as a base of prediction. The phenomenal reality of separate external objects for myself, and reputedly for other men, helps convince me that my near relatives probably use this presumptive strategy.

Tying the proximal threads in a distal knot

In the laboratory research on thing constancy, the distal thinghood of the object being judged is never in question--rather only some attribute of it--as its size, or shape, or distance. Starting from this setting one misses some of the wonder of the achievement which would be more apparent if one examined an instance in which even the positing of a single thing was equivocal. For note, there is always the alternative of inferring several independent causal sources for the several proximal data, rather than inferring a single common source (viz., Heider, 1959).

"Triangulation" is an attractive model (e.g., Feigl, 1958; Campbell, 1959). The "lens model" suggests it. From several widely separated proximal points, there is a triangulation upon the distal object, "fixing" it and its distance in a way quite impossible from a single proximal point. Binocular vision can be seen quite literally as such a triangulation.

As Brunswik emphasized, each proximal stimulation is equivocal when interpreted as evidence of a distal event. For each, there is a subinfinity of possible distal events to which it could be witness. We could if we wished add to the lens model a radiation of lines from a heterogeneity of distal events to a single proximal one. How does the cognitive apparatus decide on the distal focus, decide in what distal bundles to tie the proximal particulars? How do the two proximal sources "know" when they are fixing on one single object rather than two separate ones? Or rather, under what conditions do they "presume" one rather than two? The tentative and partial answer is that this is achieved through a pattern-matching. The tentative theme of this paper is that such a pattern matching is involved in all instances of distal knowing, including the achievements of scientific theory. In making this suggestion I join many others. For example, it is an important theme in Craik's (1943) *The Nature of Explanation*. Konrad Lorenz has made a similar point in his recent paper on "Gestalt Perception as Fundamental to Scientific Knowledge" (1959), a paper which, like the present one, emphasizes the epistemological significance of the evolution of the perceptual constancies. Bertrand Russell has been particularly explicit.

In Russell's (1948) *Human Knowledge: Its Scope and Limits* he starts out as though writing for laymen a summary of epistemological problems. And in the section on language, he seems little changed from his earlier logical atomism. But in the subsequent parts of the book he is again creatively thinking about the problems that have always troubled him. Particularly in his final section on "The Postulates of Scientific Inference" he offers a synthetic theory of inductive knowledge quite in keeping in spirit with that essayed here. While the list of implicit hypotheses about the nature of the world which he offers differs in particulars from one which I might develop, the general effort is similar.

In his "structural postulate" Russell states the principle with which the present paper deals: "When a number of structurally similar complex events are ranged about a center in regions not widely separated, it is usually the case that all belong to causal lines having their origin in an event of the same structure at the center" (p. 492). His illustrations include multiple copies of photographs, the similar percepts of people viewing a given scene or hearing a given sound, the multiple copies of a given book, the identification of which shadow goes with which man, the assumption of a common culprit in the "brides-in-the-bath" murders, etc. (Russell, 1948, pp. 460-475).

Stereoscopic vision

As a first example, let us examine binocular vision via the stereoscope. The stereoscope is ecologically a very unrepresentative sample of possible environments, and is one in which the distal-perceptual apparatus goes awry, in that two separate distal events (the separate pictures each eye views) are misinterpreted as one. In the process, the cues by which oneness is inferred are made more evident.

There are three typical outcomes, if heterogeneous pairs of pictures be allowed: Binocular fusion into a single image, double-image superimposition, and a domination of one eye's content to the exclusion of the other. If one starts out with similar simple line drawings which will fuse, this resolution can be destroyed by making one of the drawings more and more different from the other. If there is little detail in the drawings, the failure of fusion may lead to a superimposition of the two contents. For example, if each side has only two vertical lines, under the fused condition only two are seen. If the separation of the two is greatly increased beyond the capacity of binocular resolution, then three or four lines may be seen. If, on the other hand, each eye's view is rich in detail (as when two photographs are used), lack of common contour in the two pictures results in a total suppression of one or the other. It seems clear that the fusion of the two proximal sources into a distal inference is made possible through a process of pattern matching, and does not occur in the absence of a high degree of pattern similarity. Once this high degree of similarity is present to guide fusion, then minor disparities of the correct kind can produce a distal increment through the inference of the third dimension. But without an overwhelming similarity of pattern, such discrepancies cannot be utilized, as the inference of a single common source does not take place.

The above contrasts the matching of congruent patterns with the condition under which patterns are incompatible. We can draw a similar conclusion by comparing the presence of congruent patterns with the absence of any pattern. Consider the use of the stereoscope in connection with a separate reduction screen for each eye, so that punctiform view of each image is obtained. Under these conditions, no covergence takes place, even though each eye be viewing positions which would be fused if the total patterns were to be seen. Similarly, triangulation upon one of several distal light sources of heterogeneous distance and location would be impossible in a system consisting of two mechanical eyes each made up of a single photocell. Under such conditions, triangulation would be unusable just because of the unresolvable ambiguity as to whether both eyes were looking at the same source. While pattern identity can be misleading, it seems in this setting to be a minimum essential. The multiple-celledness of the retina becomes an essential requirement. No doubt it is true that the greater the degree of pattern similarity the more certain the fusion and the more inappropriate angles of optical convergence which it will overcome. From the point of view of inductive theory, the more elaborate the pattern, the more statistically unlikely a repetition of it through independent chance events becomes, and hence the more implausible the rival hypothesis of twoness becomes, in competition with the hypothesis of oneness.

Monocular processes of distal inference contain upon examination the same dependence upon pattern, and through the utilization of memory the assumption that repetitions of pattern in time come from a common source. In ~~the~~ Wertheimer's phi phenomenon of perceived movement, where two separate events are presented (under certain conditions of spatial and temporal proximity and sequence) the visual system shows bias toward a single-object interpretation so great that the hypothesis of one object in motion becomes more plausible than that of two discrete events. No doubt this effect is the stronger, and occurs over the wider spatial and temporal intervals, the more elaborate the shared pattern of the two stimuli. The recognition of visual events as similar to past ones, and hence the use of memory at all, is dependent upon pattern similarity, and is unavailable to homogeneous fragments. Similarly for patterns extended in time (Russell calls these "event structures" as distinct from the spatial patterns or "material structures," 1948, p. 464). Auditory recognition and memory are obviously dependent upon this. In animal brains at least, it seems probable that memory access or memory search is only possible on the basis of some pattern-matching resonance process (Pringle, 1951). In mechanical brains, all outcome-controlled processes including memory search are básed upon a matching process, in the older machines called "comparing relays." The detail of the pattern involved depends upon the magnitude of the alternative set within which equivocality could take place.

"Image" as a pattern of object guiding distal response

So far our focus has been on the perceptual side of distality, on human perceptual performance where verbal response effectively makes the response unproblematic. But in fact, distal responses too show imperfect mediation via equivocal proximal responses. The learned memories or instinctive guides that provide the comparison base against which motor feedback is checked must be quite multidimensional.

For this multidimensional thermostat setting, for this "input reference set" as a servosystem engineer might put it, for this blueprint of desired outcomes, or criteria against which to check the sensory inputs representing the status quo, the use of the term template seems appropriate (e.g., Campbell, 1963, pp. 143-144).

Without in any way suggesting how or in what form such a central nervous system record might be carried, it nonetheless seems logically required that there be stored templates or maps for each distal, object-consistent segment of behavior. A further problem enters here. In the servomechanism model in its simplest form, the template operates in terms of feedback after response. But such a model does not seem to be complicated enough for the typical guided distal response. For these there seem to be a perceptual search and a perceptual checking of perceptual objects against a template in terms of the discrepancy from which motor responses are guided. For organisms with vision one needs either one template against which both vision and motor feedback are compared, or else separate templates somehow coordinated. The checking of a perceptual pattern against the template (which very obviously must take place even in instinctive behavior, as in the insect recognizing a mate, or the baby chick pecking at grains of wheat) brings us closer to concepts of the storage of behavioral dispositions in terms of images. Note Uexküll's (1934) use of the concept of "search image," Tolman's (1948) "cognitive map," and Holst's and Mittelstaedt's "efference copy" (e.g., Holst, 1954; Mittelstaedt, 1962; Lorenz, 1959). Mowrer (1960), and Miller, Galanter, and Pribram (1960) have also been influenced by cybernetic considerations to reintroduce the term image. The traditional concept of schema as used by Head (1920), Bartlett (1932), and Piaget (e.g., Flavell, 1963) has a similar function.

Trial-and-error of pattern matching

In the present writer's previous contributions to a potential "psychology of knowledges processes" or "psychological epistemology," two themes have dominated. One is a perceptual theme (Campbell, 1958; 1959, pp. 172-179; 1964; Campbell & Fiske, 1959) and the other of a trial-and-error nature (Campbell, 1956; 1959, pp. 163-172; 1960b). Since many find these two themes incompatible, and since the present paper is in the perceptual vein, it seems well to emphasize that I do not at all see the points of view in disagreement, and do indeed regard both as essential. Distal knowing when involving discovery of new distal objects, or the recognition of old ones from among a set, is a trial-and-error of pattern matchings. The trial-and-error component is, of course, missing in settings in which--through other sources of knowledge--no selection is required. It is often overlooked when actually present due to the utilization of human links in a communication system, whose rapid and subtle processes are assumed but not specified.

Let us consider a possible map-guided postal missile, designed to home on a city and airport via radar pictures. Let us conceive of the system as designed to operate under conditions of ambiguity as to city, distance, and approach angle. For such a system to operate mechanically, there would have to be a trial-and-error of matchings of the current scope reading with a stored criterion map. This would have to include a mechanism for expanding and contracting the current radar picture, shifting it horizontally and vertically, tilting and rotating it, etc., comparing it with the criterion under all conditions. Some summation of discrepancy process would be involved, as through the use of a "negative" image of the criterion projected upon the "positive" of current reading, so that when the net reached a low, some guidance-locking impulse was transmitted. While the practical details are overwhelming, and might require a mechanism too large to transport, the general system is possible. In considering such a system, the blind surveying of possibilities (whether systematically, as in a radar beam, or randomly) is an essential process, only reduced insofar as already achieved knowledge, and/or reduction of potential equivocalities, is built in.

Probably for our postal missiles, such a spatial or "visual" homing system is too expensive and too uncertain. The city maps, for example, might not be distinctive enough. So the engineer would probably be tempted to employ homing on a radio beam. (The problem of contraction of pattern magnitudes with greater distance is avoided, inasmuch as the temporal "spacings" do not expand with distance from source as do the spatial ones.) If we now envisage a purely mechanical system which can "recognize" its own appropriate homing beam from among the inappropriate ones designating different systems, we see that again a pattern matching, and a trial-and-error of pattern matching is involved. Being no electrical engineer, I shall not attempt to indicate an efficient system but merely sketch out a crudely plausible one which makes explicit the problems involved. For a pattern-comparator, let us again use a subtractive procedure of combining energies, with a selector system picking out minimum energies. (Lorenz, 1959, citing Holst, has pointed to the ubiquity of these image-cancellation processes in perceptual phenomena.) Or a resonator generating a derived energy when two inputs maximally matched could be used. The process would be a "selective retention" one in that all inputs picked up would be tried out, only those which happened to resonate being "selected." It would save a lot of hardware if the total system were master-clocked so that the correct homing sender and the criterion model contained within the missile were always in perfect syncrony once the missile came in range. However, "auditory" recognition systems are not usually so restricted. To remove this restriction, a trial-and-error of lags of the criterion pattern is required, such as might be achieved by having multiple resonator units each with a recording of the criterion at a different time of onset. The received input would be blindly matched against all of these, resonance magnitude or over-all pattern similarity measures selecting the optimal one.

In creative knowledge processes involving exploring new areas of knowledge, there is not only a try and fit of extant patterns against some input, but there is also a trial-and-error of novel patterns, novel gestalten, or recenterings (Campbell, 1959b, pp. 389-390).

Pattern matching in astronomy. Astronomy, our oldest successful science, deals with the most remote and unknowable of objects. It is, therefore, an appropriate place from which to select a sample of the utilizations of pattern matching in science.

Consider first the remarkable assurance with which man assumes that he can identify "the same" star upon successive nights, even with the unaided eye. This is the more remarkable (and the more presumptive) because the stars for such instrumentation are so homogeneous a set of objects as to be mutually indistinguishable. It is further remarkable because a given star is continually changing its location, and because the average star is out of range of observation some three-fourths of the time (i.e., when below the horizon, when in ascendance during the daytime, and when eclipsed by clouds). It is the rigid pattern of the so-called "fixed" stars which make this possible, or which makes this the preferred interpretation. If indeed the stars were transient events, destroyed each dawn and constituted afresh each evening, then the recurrence across the observed ages of "the same" pattern would be a set of coincidences beyond our credulity. If there were indeed permanent stars, but these moved on a time and space scale corresponding to an enlarged Brownian movement, so that each evening a novel pattern were apparent, then astronomy would not have been the first science to emerge, nor would the confident reidentification of "the same" star on successive nights have taken place. If all the visible "stars" had been planets of our sun (a system which we now understand as simpler and more orderly than that of the fixed stars), the manifest pattern would have been so much less obvious and compelling as to have postponed the hypothesis of reidentifiable sameness for many centuries. It was the grid of the "fixed" stars that facilitated the observation of the patterned meanderings of the "wandering" stars, and enabled their reification as stable particulate substantial entities.

Astronomy's entities at this stage were diagnosed almost entirely (except for the sun and the moon) by external pattern. Each star was unidentifiable in isolation. The frame of the other stars, each in itself unidentifiable, provided the identification through its fixed pattern. In the ecology of our normal vertebrate development, most external entities are diagnosed by recurrence of internal pattern. Today with telescopes, we can so identify nearby planets, and potentially with refined spectroscopes extended into the radio frequencies we may find unique enough internal patterns to justify the hypothesis of stable thing for some of the fixed stars without the crutch of external pattern. The external pattern of the fixed stars is so strong and redundant that minor discrepancies can provide 3-D gains from successive comparisons, and through such triangulations the distances of stars and the movements of the sun itself have been inferred.

The first achievement of radar reflection from Venus (Price, et al. 1959) provides another illustration. It provides a good example because the knowledge process utilized an extremely noisy and fallible channel, and yet was so clear-cut in outcome as to make possible a correction of the prior computations of the distance of the target. Note in ordinary radar (and in television) an interchangeable transition between temporal pattern and spatial pattern, the latter being made possible by the lag in the decay of the phosphorescence on the picture tube. For ordinary uses of radar, the output beam need not be temporally patterned, but can be a constant emission in amplitude and frequency. Its figure-ground patterning comes through the contrasting reflection from object and nonobject as the antenna sweeps and re-sweeps the area. (A single punctiform and unmoving antenna's reception would be uninterpretable, due to the numerous extraneous sources of radio waves.) For the first radar reflection from Venus, this spatial patterning was not used, as not enough contrast to be visible would have resulted, the small angular size of the target and the small energy reflection being parts of the problem. Instead, a temporal pattern was imposed upon the emitting wave. The reception of a radio telescope antenna focused in the direction of the Venus was then searched for a matching pattern. The reflected signal was so weak relative to the radio frequency noise of the background that such pattern matching could only be ascertained by a cross-correlation of broadcast and reception which used two-millisecond pulses, present or absent on a fixed quasi random pattern for 4-1/2 minutes of transmission. When the cross-correlation was computed with a lag appropriate to the speed of light for the astronomically computed distance, the correlation was not above chance levels. A trial-and-error survey of shorter and longer lags located the optimal lag at a point some 5.0 milliseconds less than expected. At this lag a highly significant cross-correlation repeatedly appeared, indicating a small but significant error in the previous computation of distance. This achievement is a most impressive evidence of the power of pattern matching in identifying recurrences of "the same" thing even when the initial instructions as to "where to look" were in error.

Further notes on the ubiquity of pattern matching. The above are, of course, but a small fraction of the relevant illustrations which might be drawn of this most ubiquitous ingredient in cognitive achievements. And the present paper is, of course, only one of many appeals for the recognition of its centrality. Acknowledgment should again be made of the important paper by Lorenz (1959) on this topic. Polya (1954), Oppenheimer (1956), and Jones (1957), among others, have called attention to the important role of analogy in scientific thought. Jones has provided a bibliography on the theory of the use of models and analogues and a discussions of the concept of similarity under such conditions. The concentration of efforts on pattern recognition in the computer simulation of cognitive process is further testimony to the widespread recognition of the problem (e.g., McCulloch & Pitts, 1947; Clark & Farley, 1955; Dinneen, 1955; Selfridge, 1955; Unger, 1959; Rosenblatt, 1960; Uhr, 1963).

The pattern matching of scientific theory and data: implications for operational definitions and verification of theory

Science is the most distal form of knowing. Scientific theories are distal achievements. The processes and entities posited by science (e.g., radio stars, neutrinos, atoms, molecules, genes, cells, etc.) are all very distal objects very mediately known via processes involving highly presumptive pattern matchings at many stages. Such is a summary of the preceding pages. The present section extends this by identifying the over-all relationship between a formal scientific theory and the relevant accumulations of empirical data as one of pattern matching. The resulting interpretation is felt to be compatible with, and to summarize something common among those several modern philosophies of science which have attempted to retain the "posit" and "put-up-or-shut-up" hard-headedness of pragmatism and logical positivism, without making the error of exhaustive-definitional-operationism (Campbell, 1960a). We will but sample from these philosophies, and will not argue here their equivalence upon any grounds other than this one.

It has long been a common property among logical positivisms to describe scientific theory as an internally consistent formal logic (analytically valid) which becomes empirical (gains synthetic truth) when various terms are interpreted in a data language. A variant of this general model is accepted here. The formal theory becomes one "pattern," and against this pattern the various bodies of data are matched, in some over-all or total way. These empirical observations provide the other pattern, but somewhat asymmetrically. The data are not required to have an analytical coherence among themselves, and indeed cannot really be assembled as a total except upon the skeleton of theory. In addition, the imperfection or error of the process is ascribed to the data pattern, for any theory-data set regarded currently as "true," and except for quantum theory. Theories "known" to be in need of revision, or accepted only as convenient oversimplifications are conceptually allowed to share the residual matching error. It is as though in the radar reflection from Venus we regarded the known or intended output as the "true" pattern or theory, and the noisy reception as the data. (The asymmetrical conceptual allocation of error between experimental and dependent variables is, however, another problem, receiving explicit treatment in modern statistics.)

This variant of the "interpreted logic" version of logical positivism is in disagreement primarily with those applied varients which have taken an extreme position on "operational definitions." For the "exhaustive-definitional-operationist"if we take his "operational definitions" as defining terms in his formal theory, no error is allowed in the interpretation of theory by data. For him, the admitted imperfection of all scientific theory is located in the strength of the theoretical laws stated, i.e., within the kind of relationships which the theory posits. In such a case, it seems doubtful if a formal, analytically consistent theory is possible utilizing such terms.

While the categorization may seem to fit, Bergmann (1957) should not be identified as an "exhaustive-definitional-operationalist" in any simple manner, for note his criticisms of a similarly extreme operationism (1954) and his condoning of the reification of concepts in physics (1943). Nor should Bridgman (1927) in spite of his temptingly clear expression of this point of view in this quote: "If we have more than one set of operations we have more than one concept, and strictly there should be a separate name to correspond to each different set of operations" (p. 10) The exposition of this paper will not be hindered by regarding the "exhaustive-definitional-operationalist" as a straw man or ideal type useful in clarifying the issues but not to be identified with the position of any actual philosopher of science. While some have indirectly advocated it when arguing against the errors of other positions, the obvious weaknesses of the position as here presented have probably kept any from direct advocacy.

In contrast to definitional operationism, in the position here advocated the error in matching theory to data is allocated to the imperfect representation of the theoretical concepts by the data series. Where the measurements used by a science have a negligible proportion of error, as in macrophysics and astronomy, the difference in the points of view may be unnoticeable. In any case, focusing the difference upon a decision in the allocation of error makes it obvious that there is no analytically correct choice between the two points of view. As a description of how science has operated, I prefer the present variant, at least as the model of those segments of the physical and biological sciences which have achieved useful formal theory.

Among the logical empiricists, Feigl's critical-realistic version (e.g., 1950) is compatible with that here advocated. Hempel (1952, pp. 29-50) states the "interpreted logic" position clearly, and has attempted to preserve the values of operationism without accepting a construct-defining version of it. He calls for operational _interpretations_ of scientific terms, allowing these to be partial interpretations. His version of the operationist requirement becomes the requirement that theory be testable, i.e., interpretable so that at many points its matching with data be ascertainable. Were he to have explicitly emphasized the inherent mediateness and imperfection of all measurement processes, or to have explicitly recognized the interpretations of theory-terms as fallible, or to have explicitly located the errorfulness on the data side of the matching, his position would be indistinguishable from the one here advocated. Margenau (1950) while not a member of the logical positivist camp has a point of view regarding the relationship between theory and data which is equivalent on these points. So also are judged to be the positions of Popper (1935; 1959), Quine (1953), Hanson (1958), and Kuhn (1962) some aspects of which will be discussed more in detail below.

Note particularly that there are _two_ patterns to be compared--that of theory and that of data, even though in an iterative fashion each has developed in contact with the other. This is in disagreement with those views of science in which theory is viewed simply as a summary of data, i.e., as the simple product of inductive generalization. Popper (1959) has effectively discredited such a point of view. Hanson (1958) has emphasized this duality through his stress upon the perceptual aspects of relating theory to data, i.e., how differences in theory lead scientists to perceive the world differently, how corrections of outmoded theories must await the availability of an alternative theoretical structure. While there is a subjectivist flavor to much of Hanson's protest, descriptively his portrait of science is a pattern matching one. His discussion of Kepler's struggles to match theory and data are of particular value, as also are his detailed citations on the shifts in status of specific scientific constructs from empirical law to analytic tautology. The replacement of an old and unsatisfactory theory by a new one has in many instances the characteristic of a trial-and-error process in which total theoretical systems are tried out, being accpeted or rejected as a whole. In practice, no theory that has been judged useful in the past of a science is ever rejected simply upon the basis of its inadequacy of fit to data. Instead, it is only rejected when there is an alternative which fits better to replace it. And the fit of the new theory is not perfect, only better. Kuhn (1962) presents similar episodes in the history of science.

Of course, even for theory in physics, the above description has exaggerated the analytic (logical, mathematical) internal consistency of the theory. As Quine (1953) has pointed out, analytic systems contain hidden empirical assumptions. Where the logical or mathematical form of the theory is not complete, considerable revision of specific terms of a theory is possible without necessitating an over-all accept-reject decision. What is important is the recognition of the two-part, two-patterned nature of the process, and the acceptance or rejection of the theory or model upon the basis of some over-all criterion of fit. This [illegible] of fit is conceptualized so that any specific [illegible] may be regarded as in error, as judged by the pattern [illegible] which [illegible] be present.

is particularly clear in total-theory shifts in science (viz. Kuhn, 1962). The model of piecemeal theory revision in its extreme form makes the "theory" no theory at all, but simply a restatement of the data in its full complexity. Theory must instead be a separable pattern from data, with the fit to data problematic, otherwise testability, predictive power, and parsimony all are lost.

Fringe imperfection. It is fundamental to the general epistemology here argued that all knowledge is indirect, that all proximal stimulation is equivocal or fallible as a basis for inferring distal objects. This leads to the rejection of those epistemologies based upon any incorrigible sense data or other phenomenal givens, and a parallel rejection of a purely "proximal" science in which scientific constructs be defined in terms of (exhaustively known through) specific meter readings. Instead, all specific meters are regarded as fallible, corrigible instruments. On theoretical grounds alone, any specific meter can be seen to involve many physical laws other than the construct-relevant one. Keeping these other laws inactive through specific compensations (as in the control of inertial forces in a galvonometer) or through constant conditions is never achieved to perfection. Wilson (1952) in instance after instance provides specific illustrations of such limitations. Yet this obvious fact is apt to be negated implicitly in a definitional operationism. This negation is induced by the need to make explicit the essential relationship of scientific theory to experimental data, overdetermined by a psychic need for certainty. The pattern matching model for the fit of theory to data provides an explanation of that essential relationship which nonetheless avoids the assumption of certainty for any of the specific data. The imperfection of fit is conceptualized so that any specific meter reading can be regarded as in error, as judged by the pattern matching which minimizes error in general.

Let us consider a case in which we graph together a set of empirical points and a theoretically derived curve and achieve a good correspondence. Some of the points lie above the line, others below, but in general they fit well, and some lie "exactly" on the line. If there is no systematic deviation, we interpret the point by point deviations where they occur as error, and would expect such error to occur on some of the "perfectly fitting" points were the experiment to be replicated. While an over-all fit has been required, no single observation point has been taken as an infallible operational definition of a theoretical value--all rather are partial and fallible operational representations. Were each taken as "operational definitions," the "theory" would have to be multiple parametered enough to fit each point exactly--and if selected upon that basis one would need a new "theory" for each new set of data.

In the pattern matching of theory to data we reduce the fringe of error as much as possible, we center theory in the data points so that the fringe occurs without systematic deviation from theory, and we distribute the fringe of error over all of the observational points, potentially. We may end up saying that one observation was "right on," and that another was probably in error. A priori, however, any of the points could be wrong. It is through such a process that physicists can throw away "wild observations," an impossibility for a rigid operationist (Kruskall, 1960; Campbell, 1964). It is through such a process that physics has been able to refine its measuring instruments, a paradoxical event from the standpoint of exhaustive-definitional-operationism. Physics has at any period assumed that the great bulk of its "knowledge" was correct. From this floating platform of over-all pattern, it has then challenged and re-examined a particular measurement process. As Neurath said "We are like sailors who must rebuild their ship on the open sea" (1932). The "anchoring" of theory to data has not at all been achieved through a perfect correspondence at any particular point, but rather through a pattern matching of the two in some over-all way. The matching of the noisy radar reception from Venus with the ideal transmission pattern shows how such a pattern matching process as cross-correlation can powerfully recognize pattern while still distributing the large fringe of error over every data point. Actual statistics for estimating degree of pattern matching are, of course, not generally available, and the estimate of the human eye from graphed results is still the commonest criterion. The correlation coefficient sets a good example, however, by its equitable allocation of error.

Although Quine doubts the value of the traditional analytic-synthetic distinction which it has been convenient to employ here, he states a perspective upon the relation of theory to data quite compatible with that presented here on a number of points, but in particular, in the handling of fringe imperfection:

> "Taken collectively, science has its double dependence upon language and experience; but this duality is not significantly traceable into statements of science taken one by one. The idea of defining a symbol in use was an advance over the impossible term-by-term empiricism of Locke and Hume. The statement, rather than the term, came with Frege to be recognized as the unit accountable to an empiricist critique. But what I am now urging is that even in taking the statement as a unit we have drawn our grid too finely. The unit of empirical significance is the whole of science.

> "The totality of our so-called knowledge or beliefs, from the most casual matters of geography and history to the profoundest laws of atomic physics or even of pure mathematics and logic, is a man-made fabric which impinges on experience only along the edges. . . . A conflict with experience at the periphery occasions readjustments in the interior of the field. . . . But the total field is so undetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to re-evaluate in the light of any single contrary experience. . . . A recalcitrant experience can . . . be accommodated by any of various alternative re-evaluations in various alternative quarters of the total system, . . . but . . . our natural tendency [is] to disturb the total system as little as possible. . . . Physical objects are conceptually imported into the situation as convenient intermedia-- not by definition in terms of experience, but simply as irreducible posits. . . . Science is a continuation of common sense, and it continues the common-sense expedient of swelling ontology to simplify theory" (Quine, 1953, pp. 42-45).

Confirmation and falsification of theory. In the positivist's early effort to root out pseudo problems and metaphysics, they produced a testability criterion of meaning, so stated that confirmability of a proposition through the agreement of theory and data become its typical form. But since theories posit general laws whose verification can only spottily be sampled, and also because of the ambiguity introduced through error in data collection processes, this criterion has been challenged, particularly by Popper (1935; 1959). A strict and rigid form of it shares many of the problems of a definitional operationism, and we may regard it as an unsatisfactory statement of how experimental outcomes strengthen our belief in theory.

Popper has advocated instead a falsifiability criterion, in which a theory becomes scientifically meaningful if it is capable of being falsified by empirical data, and in which a theory becomes the better established the more experimental opportunities for falsification which it has survived and the more exacting these probes. Our established scientific theories at any time are thus those that have been repeatedly exposed to falsification, and have so far escaped being falsified. Because of its evolutionary and selective retention analogies, this criterion has appealed greatly to the present writer. As Popper explicitly recognizes (e.g., 1959, pp. 80-81), falsifiability cannot be held too strictly, as every observation refutes theory if carried out to enough decimal points. Instead, it is a selective retention of theories in competition with other theories, with the magnitude of the tolerable fringe of imperfection dependent upon the sharpness of that competition.

Note that in the pattern-matching model, the theoretical pattern is complete and continuous (with the exception of quantum theory), but the data series may be spotty and incomplete. Kepler actually had data on only a few segments of the path of Mars. If the data confirm the pattern insofar as tested, the theoretical pattern as a whole is made more tenable, including the nontested segments of the pattern. The fact that theories go beyond the as-yet-observed would not require the rejection of a confirmation criterion in a pattern-matching model, as it would in a more traditionally inductive empiricism. Further, the selective-survival of theories is now perceived to be a selection taking place in competition with other extant theories. As Popper (1959), Hanson (1958), and Kuhn (1962) all make clear, it is the absence of plausible rival hypotheses which establishes one theory as "correct." In cases where there are rival theories, it is the relative over-all goodness of fit that leads one to be preferred to another, not the absolute degree of fit of the better.

Summary

Joining analytic philosophers since Hume, ordinary and scientific knowledge of the external world has here been viewed as an analytically unjustified, highly presumptuous, and fallible process. In the dominant segments of that process, the strategy of distal knowing, of positing external objects mediately known, has been characteristic. In the several levels of distal knowing examined, pattern matching has been found to operate.

References

Bachem, A. Brain astigmatism: a discussion of space and time: *Amer. Scientist*, 1951, 40, 497-498.

Bartlett, F. C. *Remembering*. Cambridge, England: Cambridge Univer. Press, 1932.

Bergmann, G. Outline of an empiricist philosophy of physics. *Amer. J. Physics*, 1943, 11, 248-258 & 335-342.

Bergmann, G. Sense and nonsense in operationism. *Scient. Mon.*, 1954. 79, 210-214.

Bergmann, G. *Philosophy of science*. Madison: Univer. Wisconsin Press, 1957.

Bridgman, P. W. *The logic of modern physics*. New York: Macmillan, 1927.

Brunswik, E. Zur Entwicklung der Albedowahrnehmung. *Z. Psychol.*, 1928, 109, 40-115.

Brunswik, E. *Wahrnehmung und Gegenstandswelt*. Vien: Deuticke, 1934.

Brunswik, E. Psychology as a science of objective relations. *Phil. Sci.*, 1937, 4, 227-260.

Brunswik, E. The conceptual framework of psychology. *Int. Encyc. unif. Sci.*, 1, No. 10. Chicago: Univer. Chicago Press, 1952.

Brunswik, E. *Perception and the representative design of psychological experiments*. Berkeley: Univer. California Press, 1956.

Campbell, D. T. Operational delineation of "what is learned" via the transposition experiment. *Psychol. Rev.*, 1954, 61, 167-174.

Campbell, D. T. Perception as substitute trial and error. *Psychol. Rev.*, 1956, 63, 330-342.

Campbell, D. T. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. Behav. Sci., 1958, 3, 14-25.

Campbell, D. T. Methodological suggestions from a comparative psychology of knowledge processes. Inquiry, 1959, 2, 152-182.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. Amer. Psychologist, 1960, 15, 546-553. (a)

Campbell, D. T. Blind variation and selective retention in creative thought as in other knowledge processes. Psychol. Rev., 1960, 67, 380-400. (b)

Campbell, D. T. Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), Psychology: a study of a science. Vol. 6. Investigations of man as socius. New York: McGraw-Hill, 1963. Pp. 94-172.

Campbell, D. T. Distinguishing differences of perception from failures of communication in cross-cultural studies. Chapter 18 in F.S.C. Northrop & H. H. Livingston (Eds.), Cross-cultural understanding: epistemology in anthropology. New York: Harper & Row, 1964. Pp. 308-336.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. Bull., 1959, 56, 81-105.

Clark, W. A., & Farley, B. G. Generalization of pattern recognition in a self-organizing system. Proc. Western Joint Computer Conf. I.R.E., 1955, pp. 86-91.

Craik, K. J. W. *The nature of explanation*. Cambridge, England: Cambridge Univer. Press, 1943.

Dineen, G. P. Programming pattern recognition. *Proc. Western Joint Computer Conf.*, I. R. E., 1955, pp. 94-100.

Feigl, H. Existential hypotheses: realistic versus phenomenalistic interpretations. *Phil. Sci.*, 1950, 17, 35-62.

Feigl, H. The "mental" and the "physical." In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories, and the mind-body problem*. Vol. II of Minnesota studies in the philosophy of science. Minneapolis: Univer. Minnesota Press, 1958.

Flavell, J. H. *The developmental psychology of Jean Piaget*. Princeton, New Jersey: Van Nostrand, 1963.

Hanson, N. R. *Patterns of discovery*. Cambridge, England: Cambridge Univer. Press, 1958.

Head, H. *Studies in neurology*. Oxford: Oxford Univer. Press, 1920.

Heider, F. On perception and event structure and the psychological environment. *Psychol. Issues*, 1959, 1, no. 3, 1-123.

Hempel, C. G. *Fundamentals of concept formation in empirical science*. Chicago: Univer. Chicago Press, 1952.

Holst, E. von. Relations between the central nervous system and the periphery. *Brit. J. Anim. Behav.*, 1954, 2, 89-94.

Jones, R. W. Models, analogues and homologues. In G. Müller (Ed.), *Reglungstechnik: Moderne Theorien und ihre Verwendbarkeit*. München: R. Oldenbourg, 1957. Pp. 326-329.

Kuhn, T. *The structure of scientific revolutions*. Chicago: Univer. Chicago Press, 1962.

Kruskal, W. H. Some remarks on wild observations. *Technometrics*, 1960, 2, 1-3.

Loeb, J. Forced movements, tropisms, and animal conduct. Philadelphia: Lippincott, 1918.

Lorenz, K. Kant's Lehre vom apriorischen im Lichte gegenwärtiger Biologie. Blätter für Deutsche Philosophie, 1941, 15, 94-125. Translation in L. von Bertalanffy & A. Rapoport (Eds.), General Systems, Yearb. Soc. Gen. Systems Res., Vol. VII. Soc. for Gen. Systems Res. Ann Arbor, Michigan (205 N. Forest Ave.) and New York (787 United Nations Plaza), 1962. Pp. 23-35.

Lorenz, K. Gestaltwahrnehmung als Quelle wissenschaftliche Erkenntnis. Z. exp. angewandte Psychol., 1959, 6, 118-165. Translation in L. von Bertalanffy & A. Rapoport (Eds.), General Systems, Yearb. Soc. Gen. Systems Res., Vol. VII. Soc. for Gen. Systems Res. Ann Arbor, Michigan (205 N. Forest Ave.) and New York (787 United Nations Plaza), 1962. Pp. 37-56.

Margenau, H. The nature of physical reality. New York: McGraw-Hill, 1950.

McCulloch, W. S., & Pitts, W. How we know universals: the perception of auditory and visual forms. Bull. Math. Biophysics, 1947, 9, 127-147.

Miller, G. A., Galanter, E., & Pribram, K. H. Plans and the structure of behavior. New York: Holt, 1960.

Mittelstaedt, H. Control systems of orientation in insects. Annu. Rev. Entomol., 1962, 7, 177-198.

Mowrer, O. H. Learning theory and the symbolic processes. New York: Wiley, 1960.

Neurath, O., Protokollsätze, Erkenntnis, 1932-1933, 3, 204-214. Translation in A.J. Ayer Logical positivism, Glencoe: The Free Press, 1959, 199-208.

Oppenheimer, R. Analogy in science. Amer. Psychol., 1956, 11, 127-135.

Piaget, J. The child and modern physics. Scient. Amer., 1957, 196, 46-51.

Polya, G. Mathematics and plausible reasoning. Vol. I. Induction and analogy in mathematics. Vol. II. Patterns of plausible inference. Princeton: Princeton Univer. Press, 1954.

Popper, K. Logik der Forschung. Wien: Springer, 1935.

Popper, K. The logic of scientific discovery. New York: Basic Books, 1959.

Price, R., Green, P. E., Goblick, T. J., Kingston, R. H., Kraft, L. G., Pettengill, G. H., Silver, R., & Smith, W. B. Radar echoes from Venus. Science, 1959, 129, 751-753.

Pringle, J. W. S. On the parallel between learning and evolution. Behaviour, 1951, 3, 175-215.

Quine, W. V. From a logical point of view. Cambridge: Harvard Univer. Press, 1953.

Rosenblatt, F. Perceptual generalization over transformation groups. In M. C. Yovits & S. Cameron (Eds.), Self-organizing systems. New York: Pergamon, 1960.

Russell, B. Human knowledge: its scope and limits. New York: Simon and Schuster, 1948.

Selfridge, O. G. Pattern recognition and modern computers. Proc. Western Joint Computer Conf.I. R. E., 1955, pp. 91-93.

Smedslund, J. The problem of "what is learned?" Psychol. Rev., 1953, 60, 157-158.

Spencer, H. *Principles of psychology*. Vol. I., Part III, General synthesis. New York: Appleton, 1896.

Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.*, 1948, 55, 189-208.

Tolman, E. C., & Brunswik, E. The organism and the causal texture of the environment. *Psychol. Rev.*, 1935, 42, 43-77.

Uexküll, J. von. *Streifzüge durch die Umwelten von Tieren und Menschen*. Berlin: Springer, 1934. Translated in C. H. Schiller (Ed.), *Instinctive behavior*. New York: International Univer. Press, 1957.

Uhr, L. "Pattern recognition" computers as models of form perception. *Psychol. Bull.*, 1963, 60, 40-73.

Unger, S. H. Pattern detection and recognition. *Proc. I. R. E.*, 1959, 1737-1752.

Wilson, E. G. *An introduction to scientific research*. New York: McGraw Hill, 1952.